MeMAD Deliverable

5.3 Modelling Human Video Description and Best Practice Guide for Video Description

Version 1.0

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	23.06.2020
Start date of the project	01.01.2018
Due date of the deliverable	30.09.2020
Actual date of submission	08.10.2020
Lead beneficiary for the deliverable	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 theMeMAD 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

Document reviewers		
Name	Beneficiary	e-mail
Maarit Koponen	University of Helsinki	maarit.koponen@ helsinki.fi
Lauri Saarikoski	YLE	lauri.saarikoski@yle.fi

Document revisions			
Version	Date	Authors	Changes

Abstract

This deliverable draws together the three main strands of our human vs. machine descriptions research over the duration of the MeMAD project: the theoretical modelling of human engagement with multimodal texts; the comparative analysis of human and machine-generated video descriptions; and the analysis of narrative constructs and principles that human beings use in detecting and assimilating audiovisual storylines. The Deliverable focuses on **how human beings move from viewing a series of moving images containing actions and words to deriving meaning and constructing narrative, and how knowledge about human multimodal meaning-making can be drawn upon to formulate guidance for the (semi-)automation of video** captioning.

Part A of the Deliverable presents the outcomes of the narrative **modelling** of human video description, beginning with a brief overview of story grammars, which take a central part in our modelling, before presenting and discussing the findings of our empirical analyses of narrative sequencing. **Part B** presents **guidelines** for best practice in (semi-)automated video captioning. Given the prevailing challenges in this area, our guidelines focus on the generation of simple, descriptive video captions of a type most suitable for archive retrieval. They also represent a first step towards (semi-)automated methods for describing audiovisual content for audiences with additional accessibility needs.

Table of Contents

1 Introduction

The overall aim of WP5 ''*Human processing in multimodal content description and translation*" was to (a) advance our current understanding of the main principles, techniques and strategies of human-made video description by synthesising insights from previous research in Audiovisual Translation; (b) use this understanding to identify differences and commonalities of human and machine-based video description; and (c) outline human-based approaches to video description that are conducive to informing automated approaches to the description of audiovisual material.

This deliverable draws together the three main strands of our human vs. machine descriptions research delivered across WP5 over the duration of the MeMAD project: the theoretical modelling of human engagement with multimodal texts; the comparative analysis of different types of video descriptions (audio description, content description and machine description [AD, CD and MD]); and the subsequent detailed analysis of narrative constructs and principles that human beings use in detecting and assimilating audiovisual storylines. The first strand aimed to advance understanding of human approaches to video description (Deliverable 5.1), the second focused on lexico-grammatical patterns, which highlighted the shortcomings of current automated video captions, especially with regard to narrative structure (Deliverable 5.2), and the last strand's objective was modelling human-made video descriptions with a view to informing and improving automated approaches (the present Deliverable, D5.3).

Current machine learning models are applied to the automation of moving image descriptions based on deep convolutional neural networks for the purposes of visual input encoding and feature extraction (Krizhevsky, Sutskever & Hinton, 2012; Szegedy et al., 2015; He, Zhang, Ren & Sun, 2016). Recurrent neural networks, such as Long Short-Term Memory (LSTM), are used to decode these visual encodings and return a sentence that describes the multimedia content in an approximation of human captioning behaviours (Hochreiter & Schmidhuber, 1997). Although reinforcement learning (Ren, Wang, Zhang, Lv & Li, 2017), adversarial learning and adversarial inference (Park, Rohrbach, Darrell & Rohrbach, 2019) have been used to enhance the current captioning performance, the results are still broadly unreliable. Delivering moving image descriptions at a level of narrative sophistication that exceeds simple object-action labelling is therefore a major challenge. Wider availability of large-scale open access training data and improvements to the quality of human captioning in relation to these images are likely to be rewarded with more accurate results. However, sequencing descriptions into a cohesive, linear plot requires an understanding and interpretation of cues and prompts which is currently only within the scope of human beings. Consequently, one avenue which we have begun to explore in MeMAD WP5 is exploration of human approaches to narrative sequencing with a view to how these might be analysed and subsequently harnessed to inform the development of future machine learning models.

In line with this, D5.3 focuses on **how human beings move from viewing a series of moving images containing actions and words to deriving meaning and constructing narrative, and**

how knowledge about human multimodal meaning-making can be drawn upon to formulate guidance for the (semi-)automation of video captioning. Story grammars, which can be used to explain plotlines in terms of a formal structure, were tested as a way of explaining narrative sequencing from a human perspective with the intention to inform future machine learning via a broadly applicable, narrative structural framework. Our findings suggest that story grammar can be used as a means of interpreting and understanding the development of a plot in a narrative.

This deliverable is divided in two parts: In **Part A**, we present the outcomes of the narrative **modelling** of human video description, beginning with a review of progress and a brief overview of story grammars, which take a central part in our modelling, before presenting and discussing the findings of our empirical analyses of narrative sequencing.

In **Part B**, we present **guidelines** for best practice in (semi-)automated video captioning. Given the challenges that remain to be resolved in this area, our guidelines are focused on the generation of simple, descriptive video captions of a type most suitable for use in the context of archive content tagging and description, where the goal is to catalogue and retrieve material at a later date for re-use or re-sale. They also represent a first step on the long road to developing (semi-)automated methods for describing audiovisual content for audiences with additional accessibility needs, whether physical (sight-impaired) or cognitive (learning difficulties, language-related disabilities, atypical cognitive frameworks). Since audience-oriented descriptions require a far more sophisticated type of cognitive processing of the source material than content descriptions designed for film retrieval, these guidelines should be expanded in the future, in parallel with the growing sophistication of machine outputs.

PART A

2 Background: Introduction to Story Grammars

In the current phase of the project our focus has shifted away from mental modelling frameworks, which proved valuable for understanding the human interpretation of multimodal texts in the data collection and research design phases. Following quantitative assessment of the human- and machine-generated video descriptions, we looked for a way to describe and address the gap in narrative generation observed in current machine descriptions. One issue in the computer-generated video captions data we analysed was the lack of sequencing between frames and camera shots, and a failure to reflect these progressions as a developing story arc. We explored alternative theoretical frameworks to explain narrative storytelling in terms that would be transferrable to computer modelling and platform development workstreams. Story grammars appeared to fit this brief as they focus on the milestones in narrative storytelling which contribute to a sequential exposition of plot.

Story grammars came to the fore of narratological and cognitive research during the 1970s after which they fell out of favour for several decades. However, they have experienced something of a resurgence in popularity more recently. Originally featuring in studies related to storytelling for pedagogy (Singer & Donlan, 1982), and in the field of narrative recall (Rumelhart & Ortony, 1977; Stein & Glenn, 1979) story grammars were seen as a way of describing and systematising the stages of dramatic exposition in cohesive narrative arcs. Currently, there is interest in using story grammars *inter alia* to understand macro-narrative in children's storytelling (Appose & Karuppali, 2018), to coach cognitively challenged children to recount witness statements when giving evidence in court (Brown, Brown, Lewis & Lamb, 2018), and to investigate deficiencies of global cohesion in relaying narrative amongst autistic spectrum children (Banney, Harper-Hill & Arnott, 2015; Whalon, Henning, Jackson & Intepe-Tingir, 2019). All of these applications lend credibility to the application of story grammars to explain and systematize human storytelling in its many forms.

The value of story grammars to this study is that they represent a shorthand for marking progression and narrative shifts across a story arc. They lend themselves to the disassembly and classification of narrative into component parts. Foremost in the development of story grammars were Rumelhart (1975, 1980), Stein & Nezworski (1978), Nezworski, Stein & Trabasso (1982), Mandler & Johnson (1977, 1980) and Lehnert (1981). At one extreme, Rumelhart's early contribution to narrative analysis comprised a simplistic formula for categorizing stories into three parts (effectively, beginning, middle and end). Lehnert (1981), at the other extreme, elaborated a concept of 'plot units' with algebraic levels of complexity and hyper-granular sub-plot categories, assigning units labels like 'positive trade off', 'motivation', 'perseverance' and 'mixed blessing'.

In the early stages of our investigation we explored Lehnert's schema at some length. While acknowledging the worth of a detailed bottom-up approach to narrative analysis, we found that the assignment of unit labels was highly subjective in practice, and ultimately the level

of information retrieved was too granular to be helpful at this stage of automatic video description development. By contrast, Rumelhart's (1975) early schema operates at a more fundamental level, dividing a narrative into three principle components: *event*, *goal* and *outcome*. Its value, but also its limitation, is that this schema is sufficiently generic to apply to almost any type of narrative, working equally well for a news story or documentary as it does for a film extract or fairytale. Indeed, this is precisely Rumelhart's point: "story grammars have been useful in determining relevant portions of a story as a basis for a theory of summarizing, as a generally applicable scheme for analyzing a wide range of stories" (1980:316). With the exception of the most avant-garde of narratives it is generally possible for human beings to identify a dramatic *event* within any plot that drives the protagonists towards a desired *goal*, and through the realization of, or failure to achieve, that goal produce a narratively salient *outcome.* The value to be gained from this simple trifecta approach to segmentation is that we are able to detect, in the most fundamental terms, the forward trajectory of a plot towards the finish line.

Up until this point, artificial intelligence and machine learning have not become sufficiently elaborate to perform this simple act of detecting narrative progression without human intervention. Computer models still treat each movie frame as a new item of data without cohesive ties to the past or future plotline. Rumelhart's prescient view of story grammars, and the potential of story schemata, foreshadowed his own work in the field of machine learning. Re-visiting this research has proved useful in framing our own results, but alongside Rumelhart's perspective we looked for an alternative rubric that was less granular than Lehnert's work, yet more descriptive than Rumelhart's approach.

Stein and Nezworski (1978) devised a more comprehensive method of story segmentation, creating a grammar which described "the higher order structures regulating the organization and retrieval of incoming story information" (1978:177). In doing so, they sought an explanation for the kind of 'logical relations' which joined together 'story components', for the principal purpose of memory and recall (1978:177). Their schema comprised six components: *setting, initiating event, internal response, attempt, consequence* and *reaction* (1978:182). Mandler and Johnson's (1977, 1980) work ran in parallel with that of Stein and Nezworski (1978), with the former adopting segmentation labels termed 'nodes' (*setting, beginning, reaction, attempt, outcome, ending*), which were characterized by a marked fluidity in application (for instance, repetition of labels was permitted). Acknowledging the narrative value of their work on a mental modelling level, Mandler and Johnson stated, "in addition to content specifications, stories have specifiable structures and […] people have knowledge about such structures which they use in the course of comprehension and retrieval" (1980:311). It is this structural knowledge that the machine must be trained to recognise.

One of the questions in determining whether story grammars can be useful for explaining narrative storytelling in a way that would be transferrable to computer modelling, is to what extent SG schemata can be applied to multimodal/audiovisual material. We applied SG

analysis on a representative sample of our MeMAD500 video corpus to test the hypothesis that human beings make sense of multimodal texts by superimposing a pre-existing schema in order to contextualise and 'decode' narratively linked events. Our approach to analysing the applicability of SG is outlined in section 3.

3 Methodology

3.1 Aim of the study and overall approach

In line with the question raised above, the overall aim of our SG-related work is to explore whether there is value to be found in applying SG to the identification of narrative units and plot lines in multimodal material, in order to determine the steps necessary to train a computer in the art of image sequencing and narrative exposition. As a first step towards this aim, we explored whether the validity of SG as a schema in human processing of audiovisual narrative can demonstrated through a process of analysing and segmenting audiovisual material, by using the main SG elements. Specifically, our objective was to ascertain

(a) The extent to which the SG elements can be used to segment selected film extracts from the MeMAD500 video corpus;
(b) The contribution that different types of cue (visual, audio) make to segmentation decisions;
(c) The level of agreement that exists between the annotators with regard to (a) and (b).

Our assumption, consistent with the insights from mental model theory, which our earlier work in WP5 highlighted, was that viewers construct meaning within micro- and macro-scale narratives by seeking form within the sequencing, and that there is a learnt behaviour in doing so. In its simplest form, a narrative contains a *beginning*, dramatically *significant event* and associated climax, and *end* or resolution. This is quite different from stating that the plotline consists of a *start, middle and finish*. The former denotes narrative significance, with a commencement of actions and events which are salient to plot exposition, a narratively significant 'centrepiece' accompanied by a dramatic climax, and a conclusion which renders consequence to the narratively relevant event; whereas the latter denotes temporal milestones, signifying the commencement, duration and cessation of a multimodal text.

To explore the applicability of the SG approach, we initially applied the most fundamental of story grammars in our analysis. Rumelhart's schema (*event, goal, outcome*) could be readily assigned to all extracts with a seemingly good fit. Its generic, simplistic nature offers itself readily to adoption as a means of training computer models to detect the general shape of narrative. However, it fails as an explanation for the subtle stages in narrative development which fall within each of these broader categories and are necessary in order to train the computer to 'interpret' a narrative arc. For this reason, while using Rumelhart's schema as a starting point, we selected Stein and Nezworski's (1978) more explicit SG schema as the focus of our attention for annotating and analysing plot segmentation within our own corpus of film extracts. The application of both schemes suggested that Rumelhart's *event* tended to correspond to Stein and Nezworski's (1978) *setting, initiating event* and *internal response*, his *goal* matched their *attempt*, and *outcome* followed the same narrative template as the *consequence* and *reaction* segments (Stein and Nezworski, 1978) notation.

3.2 Selection of film extracts for SG analysis

The initial step was to choose twenty film extracts from the MeMAD500 corpus of short video extracts developed in WP5. The extracts were selected for their relatively well developed story arcs and reflect the diversity of genres present in the MeMAD500 corpus. The extracts were selected from three categories: extracts including dialogue (10), extracts including monologue (5) and extracts without speech (5). Dialogue extracts were used to explore whether conversational pragmatics influenced SG segmentation; nil speech extracts allowed us to explore visual cues in isolation from speech, and determine whether they alone convey narrative segmentation cues. It needs to be acknowledged that although the latter category contained no speech, sound effects such as background music, dogs barking, etc. were occasionally present and so contributed to the construct of narrative. In terms of duration, the shortest extract was 24 seconds long and the longest was 1 minute and 40 seconds.

3.3 Segmentation and annotation procedure

Our study involved two annotators i.e. the project researchers, who each analysed and segmented the selected film extracts independently, using the SG elements/labels as the basis for the segmentation. This enabled us to examine commonalities and discrepancies between the two annotators' decisions, to identify the SG elements with the highest/lowest discrepancies and to explore whether the nature of these discrepancies was conceptual or whether it was caused by the fuzziness of the segment boundaries.

The annotators used the ELAN software package, which supports multi-level (tier-based) annotation of audio and video recordings (Wittenburg, *et al.* 2006). The annotators identified key markers indicating shifts in the narrative and, based on this, divided each extract into segments based on the two SG schemata we used in the study, i.e. Rummelhart's and Stein and Nezworski's schemata (see Figure 3.1 below). The annotators then also analysed audio/visual cues occurring at the segment boundaries with a view to the role they play and the weight they have in segmentation decisions.

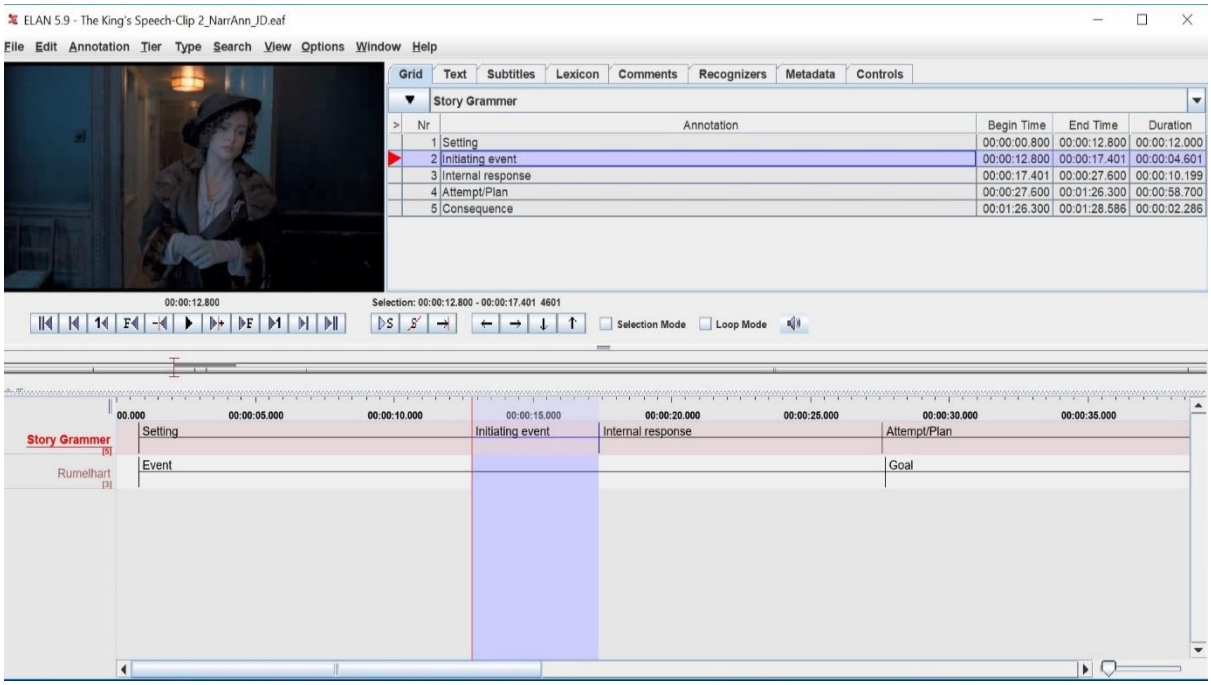

Figure 3.1: ELAN interface

After completing the segmentation and analysis, the annotators discussed their decisions with the aim of rationalising their respective segmentations and their perceptions of the role of different audio and visual cues in the segmentation process. The outcomes of this discussion are noted in the findings alongside the quantitative analysis of the discrepancies.

3.4 Data processing and analysis

After completion of the segmentation and analysis of audio and visual cues, the time-in and time-out values of each annotator's segments for each film extract were transferred to a spreadsheet, and the duration of each segment was calculated. Subsequently, the discrepancies between the two annotators with regard to the time-in points, time-out points and segment durations were calculated. In addition, the annotators recorded for each segment whether the main cue guiding them in setting the segment boundaries was a visual or audio cue. Table 3.1 shows the values recorded for one of the 20 extracts.

SG element	Ann1 (A1)				Ann2 (A2)				Discrepancy: A1 vs A2		
	Time in	Time out	Duration	Cue	Time in	Time out	Duration	Cue	Time in	Time out	Duration
Setting	00:00:00.000	00:00:04.430	00:00:04.430	V	00:00:00.000	00:00:02.990	00:00:02.990	V	00:00:00.000	00:00:01.440	00:00:01.440
Initiating event	00:00:04.430	00:00:10.780	00:00:06.350	V	00:00:02.990	00:00:06.340	00:00:03.350	A	00:00:01.440	00:00:04.440	00:00:03.000
Internal response	00:00:10.780	00:00:13.999	00:00:03.219	V	00:00:06.340	00:00:09.490	00:00:03.150	V	00:00:04.440	00:00:04.509	00:00:00.069
Attempt	00:00:13.999	00:00:27.675	00:00:13.676	V	00:00:09.490	00:00:11.452	00:00:01.962	A	00:00:04.509	00:00:16.223	00:00:11.714
Consequence	00:00:27.675	00:00:41.138	00:00:13.463	A	00:00:11.452	00:00:27.630	00:00:16.178	V	00:00:16.223	00:00:13.508	00:00:02.715
Reaction	00:00:41.138	00:00:43.459	00:00:02.321	V	00:00:27.630	00:00:43.669	00:00:16.039	A	00:00:13.508	00:00:00.210	00:00:13.718

Table 3.1: Timings for extract 103312 [The Aviator]

4 Findings

4.1 Segmentation: overview

As a first step in our analysis, to explore whether the SG schemata can be applied to the analysis of our audiovisual micro-narratives, the data relating to each SG element were garnered from across the 20 film extracts, and analysed with regard to discrepancies between the two annotators in the time in, time out and duration. Table 4.1 shows the discrepancies in duration for *setting*.[1]

Extract number	Extract title [clip ID in MeMD500 corpus]	Discrepancy in duration	Discrepancy above mean
1	Extremely Loud and Incredibly Close_ [100900]	00:00:16.360	x
2	Johnny English Reborn_[201500]	00:00:26.901	x
3	The Aviator_ [103312]	00:00:01.440	
4	The Devil Wears Prada_ [203409]	00:00:00.710	
5	The Guardian_ [203614]	00:00:00.688	
6	The Help_ [203701]	n.a.	
7	The Social Network_ [004303]	00:00:00.125	
8	An Education_[000200]	00:00:00.708	
9	Sex and the City_[102706]	00:00:01.578	
10	The King's Speech_[004001]	00:00:01.942	
11	Goal_ [201313]	00:00:01.386	
12	Memoirs of a Geisha_ [101803]	00:00:05.200	
13	Click_[000703]	00:00:00.250	
14	Bruce Almighty_ [000501]	00:00:00.162	
15	Being Julia_[200310]	00:00:08.240	x
16	The Aviator_ [103311]	00:00:13.442	x
17	Pretty Woman_[102404]	00:00:10.586	x
18	The Matador_ [004100]	n.a.	
19	The Forgotton_[003502]	00:00:03.705	
20	Casino Royale_[200609]	00:00:00.568	
	Mean	**00:00:05.222**	

Table 4.1 The average discrepancy for 'Setting'

The same process was carried out for each SG element to compute mean discrepancies for each SG element (Table 4.2). Extracts with an above-average discrepancy between the two annotators in one or more SG elements were noted for further investigation, especially to explore whether the scene type and the level of complexity of the narrative might be the cause for a higher level of disagreement between the two annotators.

SG element	Mean discrepancy	Std dev	Extracts with duration discrepancy above mean
Setting	00:00:05.222	00:00:07.126	Extremely Loud…, Johnny English Reborn, Being Julia, The Aviator [103311], Pretty Woman
Initiating event	00:00:05.356	00:00:05.124	Extremely Loud…, The Devil Wears Prada, Social Network, The King's Speech, Bruce Almighty, Being Julia, The Matador, The Forgotten
Internal response	00:00:04.377	00:00:05.948	Extremely Loud… , The King's Speech, The Matador, The Forgotten
Attempt	00:00:13.581	00:00:16.082	Extremely Loud…, The Aviator [103312], Sex and the City, The King's Speech, The Aviator [103311], Casino Royale
Consequence	00:00:06.795	00:00:08.118	Extremely Loud…, The Devil Wears Prada, The Forgotten, Casino Royale
Reaction	00:00:05.332	00:00:06.796	Extremely Loud…, Johnny English Reborn, The Aviator [103312], The Help, An Education, The Aviator [103311]

Table 4.2 The average discrepancy for all SG elements in twenty selected extracts

[1] No values were assigned to two extracts, because one Annotator merged the *setting* and *initiating* elements in the segmentation of these extracts, believing that this would be the best way of representing the narrative exposition. Similarly, where the same segment label was used multiple times across one extract (for instance, where two *attempt* segments were identified), the segments with the same label were grouped together to achieve a comparison between annotators.

The **lowest** discrepancies arose for *internal response*, *setting*, *reaction* and *initiating event*. Whilst inter-annotator agreement was expected for the beginning (*setting*) and the ending (*reaction*) of the narrative, the data suggest that the *internal response* is the narrative milestone with the highest agreement between the two annotators. This is an interesting observation, which should be explored with further participants to confirm (or otherwise).

The **highest** discrepancy between the annotators (more than 13 seconds) arose with regard to the *attempt* element, which presents the part of the narrative focussed on achieving the protagonist's goal. Several factors may have contributed to this. Most important, the interpretation of a protagonist's move as an *attempt* can, for instance, be brought about by the protagonist's physical action to demonstrate their attempt, by an audio cue, or by an abstract concept such as a thought or a plan the human viewer has in mind and uses to make sense of the move. To illustrate this point, Figure 4.1 and Figure 4.2 below show the annotators' creation of the *attempt* segment for the *Casino Royale* extract. In this extract, James Bond arrives in his suite and notices a broken wine glass on the table. He squints, hears water running in the shower and walks towards it. He pushes the door open and sees Vesper (the Bond girl), sitting in the shower in her evening dress, her body trembling and looking scared.

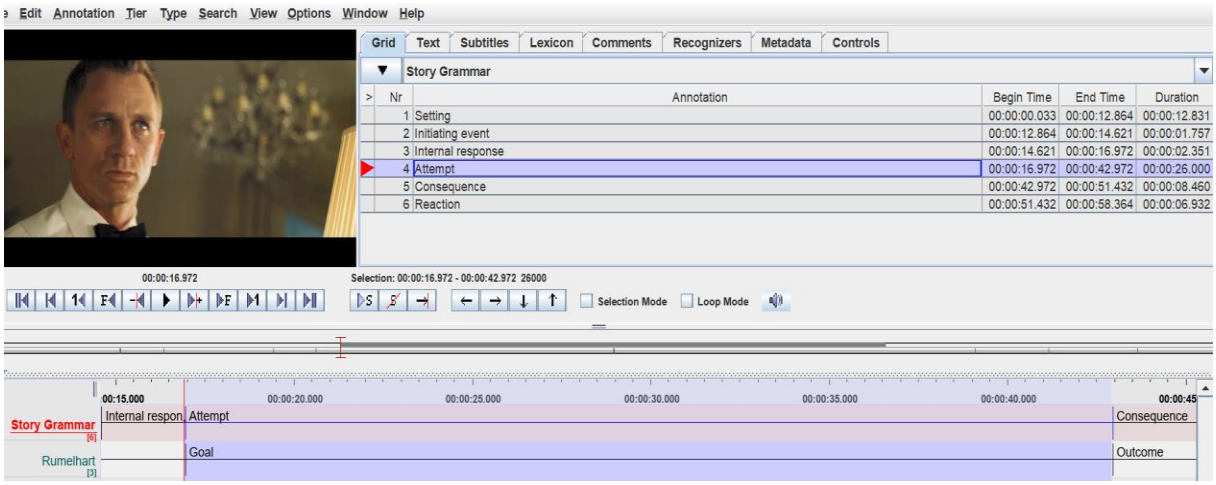

Figure 4.1: Segmentation of the Attempt by Ann1

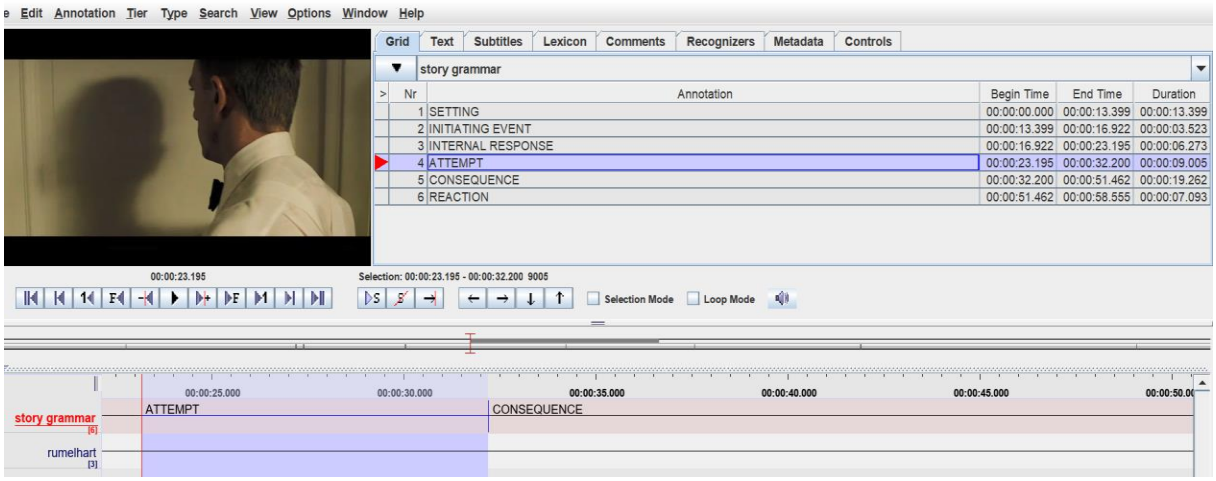

Figure 4.2 Segmentation of the Attempt by Ann2

Ann1's time-in for this segment is 00:00:16.972, and the segment's duration is 26 seconds. For this annotator, the protagonist's *attempt* begins when the protagonist squints after noticing the broken glass (visual cue) and moves towards the bathroom to find out why the shower is running (as he had not expected anyone else to be in the suite); whereas, for Ann2, the *attempt* begins when the protagonist is pushing the shower door open (00:00:23.195; also a visual cue). Figure 4.3 shows both annotators' time-in and time-out points for the attempt element. Both interpretations make sense and are equally valid, highlighting that human recipients can perceive and process subtleties in a narrative in different ways based on factors including common knowledge and individual experience.

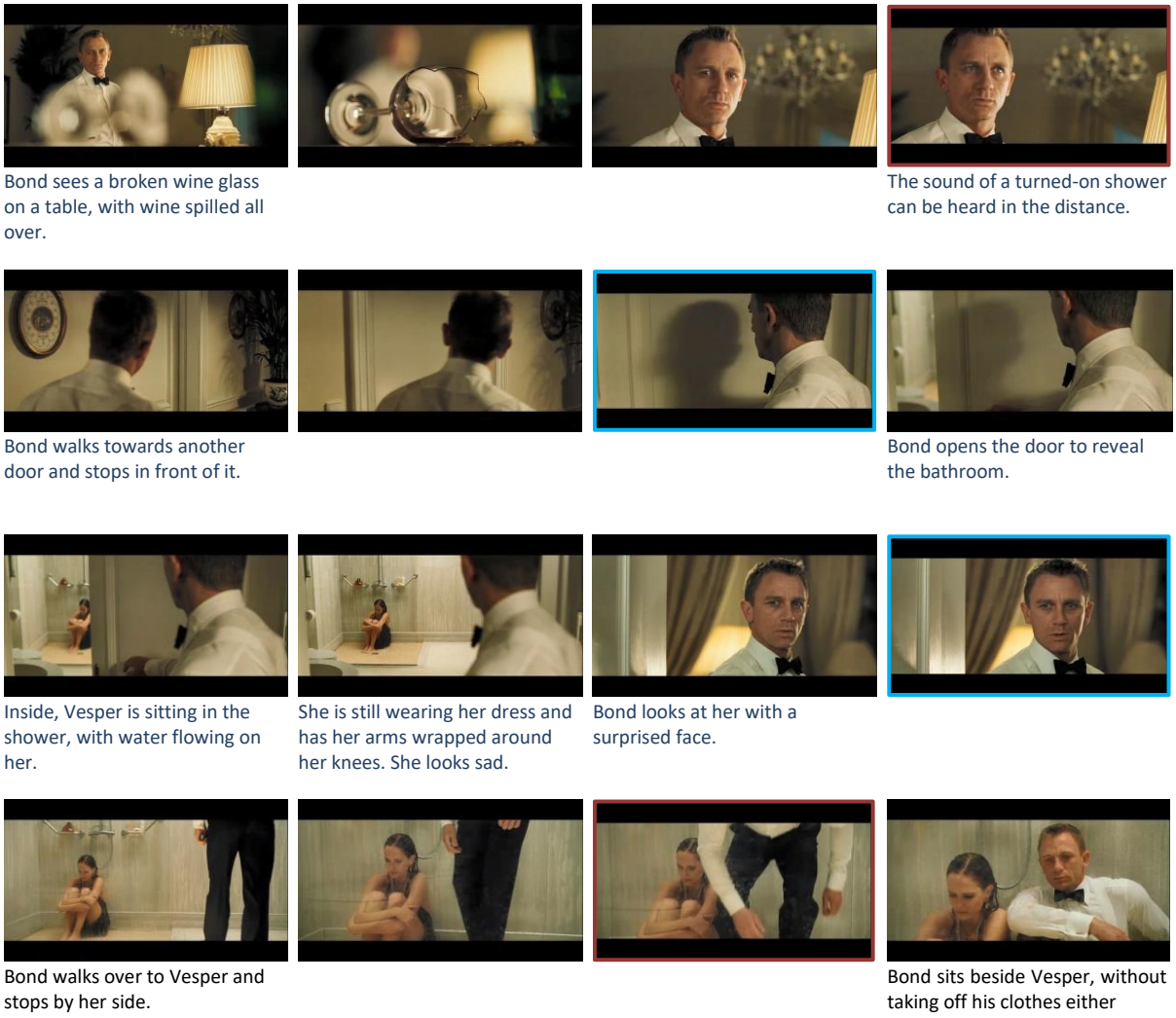

Figure 4.3 Casino Royale_ 'attempt' (Ann1 (Purple frames): time in: 00:00:16.972, time out: 00:00:42.972; Ann2 (Blue frames): time in 00:00:23.195, time out: 00:00:32.200)

This example illustrates disagreement with regard to both beginning/end and duration of the segment. This is the case for most segments, but as mentioned above, an important question is whether the discrepancies arise from conceptual differences in the segmentation or whether they are a result of fuzzy segment boundaries in audiovisual narrative, without necessarily representing a meaningful difference. This will be explored in the next section.

In order to establish consistency of protocols between annotators, discussions were held before the annotation process started to ensure work methods for determining segmentation labels and their application in the ELAN timeline were uniform. Any queries arising in relation to the application of protocols during the segmentation process were discussed in general terms (i.e. avoiding reference to a specific extract). Annotators used a reference sheet containing Stein & Nezworski's (1975) segmentation definitions as a prompt to ensure these were also applied in a uniform manner.

4.2 Segmentation: analysis of individual SG elements

The *duration discrepancies* in relation to each SG segment were transferred into scatter charts to demonstrate the distribution of the discrepancies against the average for the SG element.

4.2.1 Setting

Setting marks the opening of the narrative, or according to Stein & Nezworski, the introduction of the protagonist, which might also contain other criteria such as "social, physical, or temporal context" of the rest of the story (1978:178). The duration discrepancy between the two annotators was expected to be low, particularly with regard to the time-in for this segment. The findings support this expectation to some extent, as *setting* had the second-lowest average discrepancy amongst the SG segments (M=5.222; SD = 7.126), with seven of the 20 film extracts showing discrepancies of less than one a second (extracts 4, 5, 7, 8, 13, 14 and 20), and thirteen extracts falling below mean.

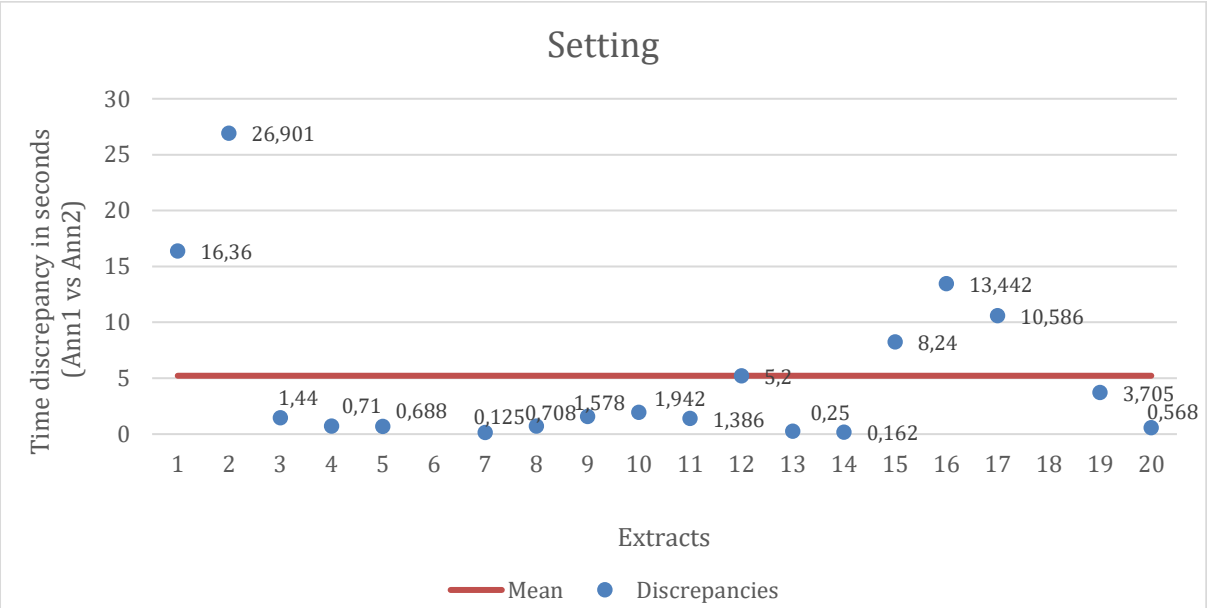

Figure 4.4: Duration time discrepancy between Ann1 & Ann2 for Setting (Mean: 5.222, Standard deviation: 7.126)

No	Extract title	Extract ID	Discrepancy (in seconds)	No	Extract title	Extract ID	Discrepancy (in seconds)
1	Extremely Loud…	[100900]	16.360	11	Goal	[201313]	1.386
2	Johnny English Reborn	[201500]	26.901	12	Memoirs of a Geisha	[101803]	5.200
3	The Aviator	[103312]	1.440	13	Click	[000703]	0.250
4	The Devil Wears Prada	[203409]	0.710	14	Bruce Almighty	[000501]	0.162
5	The Guardian	[203614]	0.688	15	Being Julia	[200310]	8.240
6	The Help	[203701]	N/A	16	The Aviator	[103311]	13.442
7	The Social Network	[004303]	0.125	17	Pretty Woman	[102404]	10.586
8	An Education	[000200]	0.708	18	The Matador	[004100]	N/A
9	Sex and the City	[102706]	1.578	19	The Forgotten	[003502]	3.705
10	The King's Speech	[004001]	1.942	20	Casino Royale	[200609]	0.568

Table 4.3 Extract IDs with the discrepancies for' Setting'

However, there are some larger discrepancies, e.g. for extract 2, *Johnny English* Reborn (26.901 seconds). This extract begins with two protagonists introducing themselves to one another. For Ann1, the end of the introduction constituted the end of *setting* (second 5.480), and the subsequent action by one protagonist offering a seat to the other was the *initiating* event, whereas, for Ann2, *setting* was considerably longer and included the introduction, offering a seat and starting a conversation. In Ann2's perception, the *initiating* event began when wind starts blowing through the window, scattering the papers all over the place.

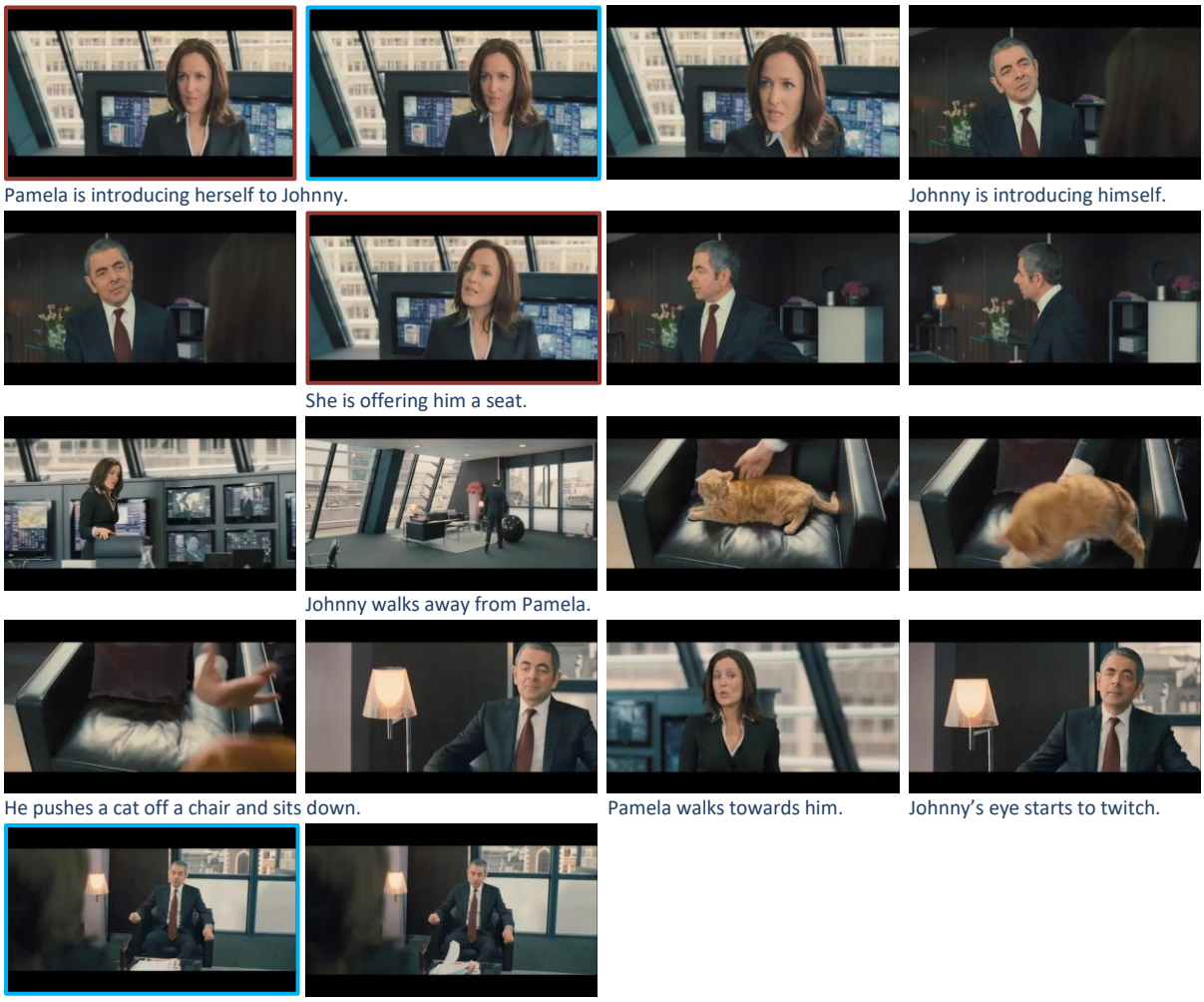

Figure 4.5. Johnny English Reborn_ 'Setting' (Ann1(Purple frames): time in: 00:00:00, time out: 00:00:05.480; Ann2 (Blue frames): time in: 00:00:03, time out: 00:00:32.411)

4.2.2 Initiating event

Initiating event "marks a change in the story environment" (Stein & Nezworski, 1978:178). In other words, "an action, an internal event, or a natural occurrence which serves to initiate or to cause a response in the protagonist" (1978:182). The annotators' interpretation of *initiating event* varied somewhat as each individual saw a change in the "story environment" in a different way.

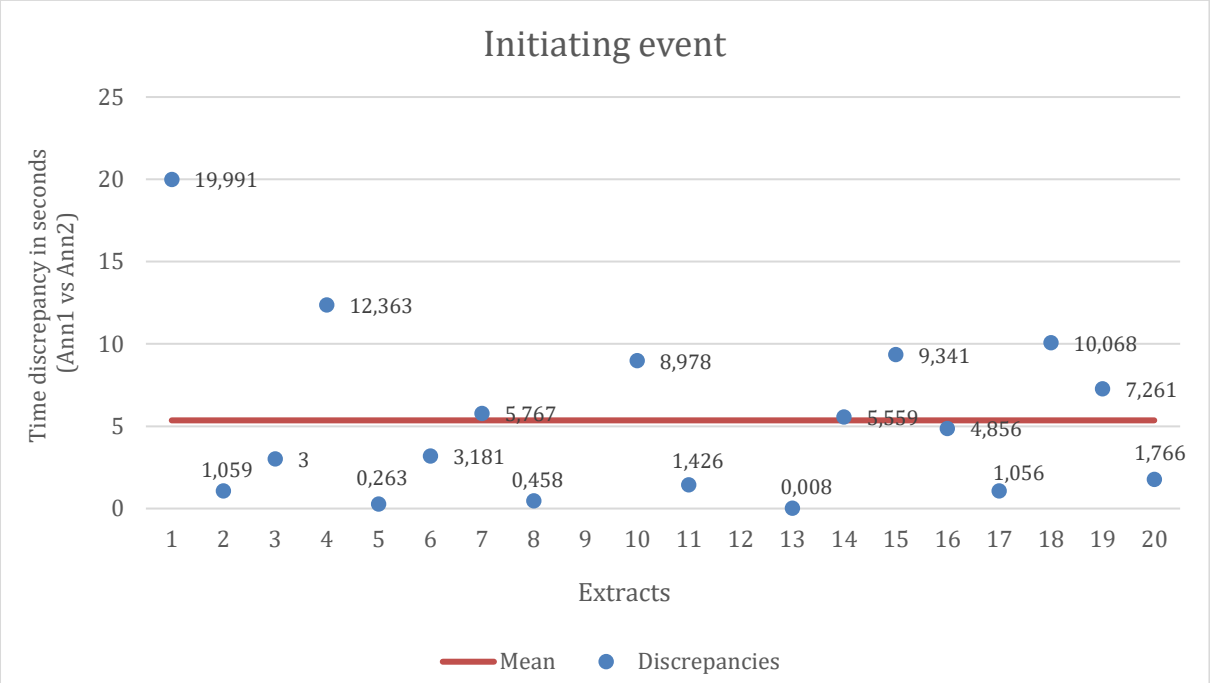

Figure 4.6 Duration time discrepancy between Ann1 & Ann2 for 'Initiating Event' (Mean: 5.356, Standard deviation: 5.124)

No	Extract title	Extract ID	Discrepancy	No	Extract title	Extract ID	Discrepancy
1	Extremely Loud …	[100900]	19.991	11	Goal	[201313]	1.426
2	Johnny English Reborn	[201500]	1.059	12	Memoirs of a Geisha	[101803]	N/A
3	The Aviator	[103312]	3.000	13	Click	[000703]	0.008
4	The Devil Wears Prada	[203409]	12.363	14	Bruce Almighty	[000501]	5.559
5	The Guardian	[203614]	0.263	15	Being Julia	[200310]	9.341
6	The Help	[203701]	3.181	16	The Aviator	[103311]	4.856
7	The Social Network	[004303]	5.767	17	Pretty Woman	[102404]	1.056
8	An Education	[000200]	0.458	18	The Matador	[004100]	10.068
9	Sex and the City	[102706]	N/A	19	The Forgotten	[003502]	7.261
10	The King's Speech	[004001]	8.978	20	Casino Royale	[200609]	1.766

Table 4.4 Extract IDs with the relating discrepancies for 'Initiating Event'

For instance, in extract 1 (*Extremely Loud and Incredibly Close*), the *initiating event* for Ann1 begins when the main protagonist (a nine-year old boy) notices something is odd about two people he is talking to, whereas for Ann2 it beings when the boy starts to explain the reason why he is there (16 seconds later than Ann1's time-in). Another interesting instance for this segment was observed in one of the nil-speech extract, namely extract 17 (*Pretty Woman*). Ann1 considers the moment the protagonist notices a blond wig as the *initiating event,* whereas Ann2 marked the very beginning of the extract when the man is deep in thoughts under the shower as the *initiating event* (Figure 4.7). As with *setting*, both interpretations represent valid points of view. Although interpretations of the change in story environment

vary somewhat between the annotators, the discrepancies are above the mean for only six of the twenty extracts and only some of these six, including extracts 1 and 17, are based on different conceptualisations of the *initiating event*.

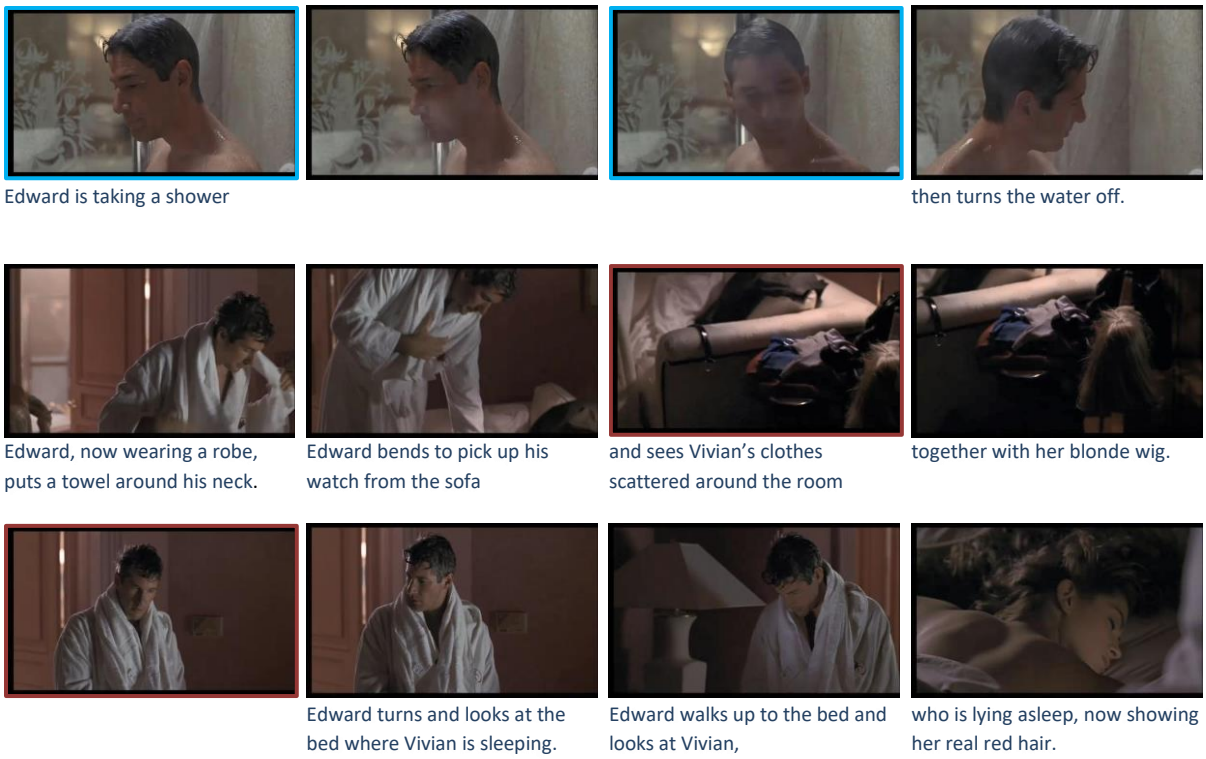

Figure 4.7 Pretty Woman_ 'Initiating event' (Ann1 (Purple frames): time in: 00:00:15.499, time out: 00:00:18.415; Ann2 (Blue frames): time in: 00:00:04.913, time out: 00:00:06.773)

4.2.3 Internal response

This SG element is defined as the response evoked from the protagonist by the *initiating event*. It can take the form of "an emotion, cognition, or goal of the protagonist" (Stein & Nezworski, 1978:182). However, one observation we made is that an *internal response* can be observed either in the protagonist initiating an event, and/or in a second party.

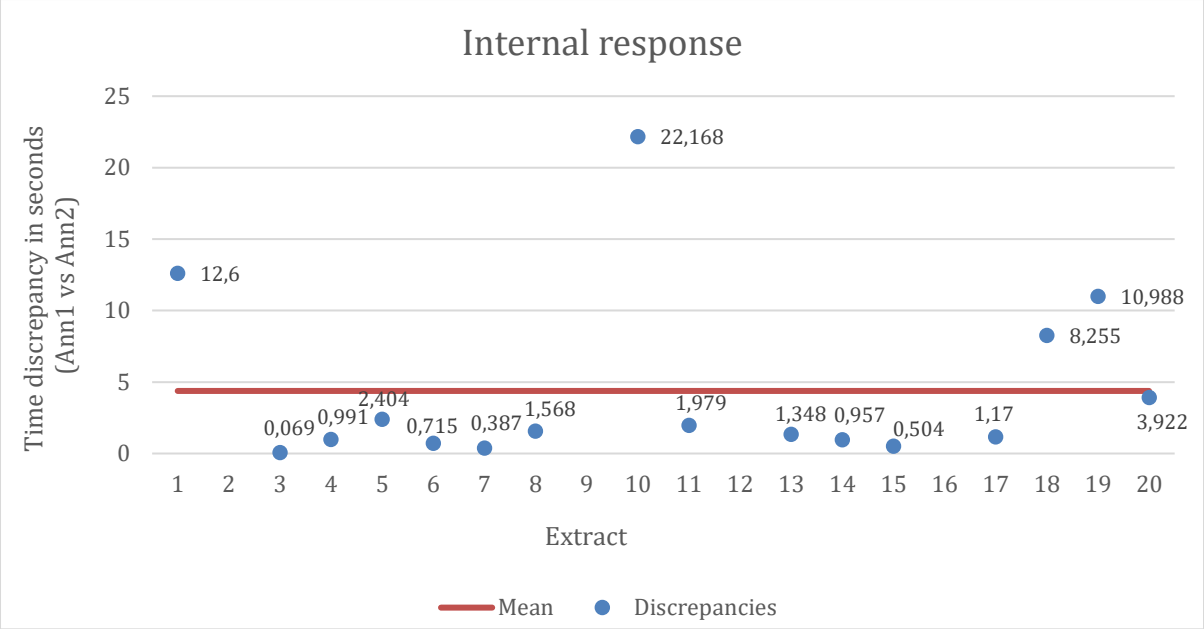

Figure 4.8 Duration time discrepancy between Ann1 & Ann2 for 'Internal Response' (Mean: 4.377, Standard deviation: 5.948)

No	Extract title	Extract ID	Discrepancy	No	Extract title	Extract ID	Discrepancy
1	Extremely Loud …	[100900]	12.6	11	Goal	[201313]	1.979
2	Johnny English Reborn	[201500]	N/A	12	Memoirs of a Geisha	[101803]	N/A
3	The Aviator	[103312]	0.069	13	Click	[000703]	1.348
4	The Devil Wears Prada	[203409]	0.991	14	Bruce Almighty	[000501]	0.957
5	The Guardian	[203614]	2.404	15	Being Julia	[200310]	0.504
6	The Help	[203701]	0.715	16	The Aviator	[103311]	N/A
7	The Social Network	[004303]	0.387	17	Pretty Woman	[102404]	1.17
8	An Education	[000200]	1.568	18	The Matador	[004100]	8.255
9	Sex and the City	[102706]	N/A	19	The Forgotten	[003502]	10.988
10	The King's Speech	[004001]	22.168	20	Casino Royale	[200609]	3.922

Table 4.5 Extract IDs with the relating discrepancies for 'Internal Response'

As a case in point, in extract 9 (*Sex and the City*), both protagonists express narratively related *internal responses a*nd *reactions*. In other words, the story arc can be seen through two separate narratives in one single extract and not merely with a focus on one character, as the protagonists can cause and influence internal responses and reactions in one another. It is perhaps one of the shortcomings of SG that it does not appear to provide enough flexibility for such rather complex extracts. However, this SG element had the lowest mean discrepancy in segment duration (4.377 seconds), with only four extracts above mean.

4.2.4 Attempt

As was briefly explained earlier in this report, *Attempt* is defined as "an overt action to obtain the protagonist's goal" (Stein & Nezworski, 1978:182). This element has the highest average discrepancy (M=13.581, SD 16.082). Nonetheless, only five extracts have average discrepancies above mean, suggesting the discrepancies for this SG element are few in number but greater in duration.

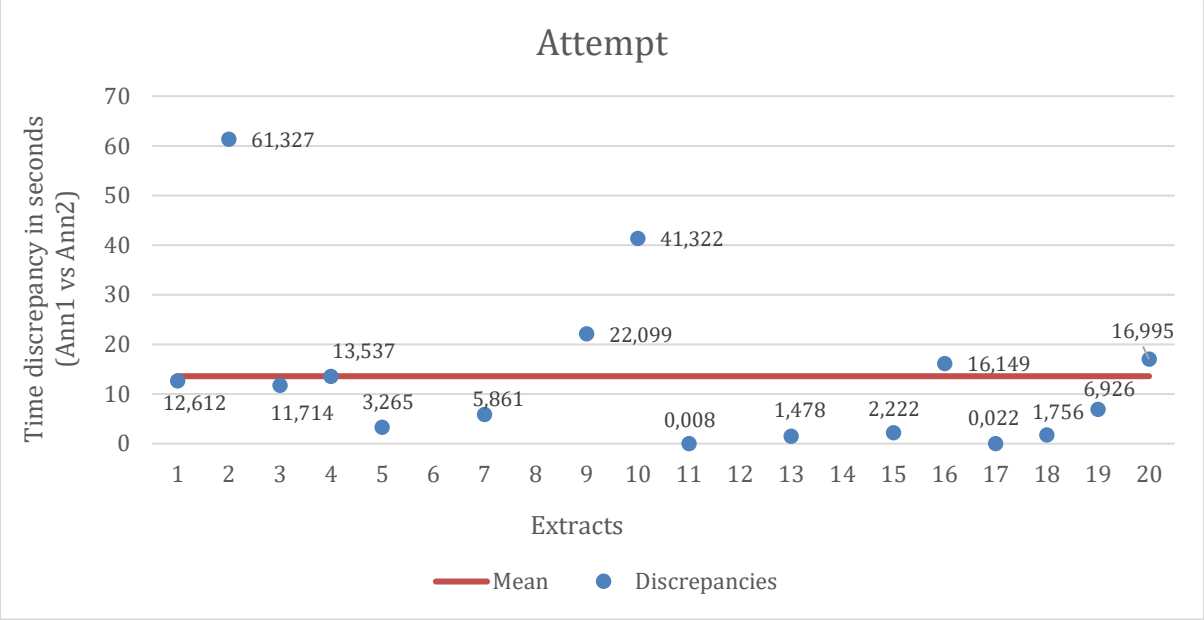

Figure 4.9 Duration time discrepancy between Ann1 & Ann2 for 'Attempt' (Mean: 13.581, Standard deviation: 16.082)

No	Extract title	Extract ID	Discrepancy	No	Extract title	Extract ID	Discrepancy
1	Extremely Loud …	[100900]	12.612	11	Goal	[201313]	0.008
2	Johnny English Reborn	[201500]	61.327	12	Memoirs of a Geisha	[101803]	N/A
3	The Aviator	[103312]	11.714	13	Click	[000703]	1.478
4	The Devil Wears Prada	[203409]	13.537	14	Bruce Almighty	[000501]	N/A
5	The Guardian	[203614]	3.265	15	Being Julia	[200310]	2.222
6	The Help	[203701]	N/A	16	The Aviator	[103311]	16.149
7	The Social Network	[004303]	5.861	17	Pretty Woman	[102404]	0.022
8	An Education	[000200]	N/A	18	The Matador	[004100]	1.756
9	Sex and the City	[102706]	22.099	19	The Forgotten	[003502]	6.926
10	The King's Speech	[004001]	41.322	20	Casino Royale	[200609]	16.995

Table 4.6 Extract IDs with the relating discrepancies for 'Attempt'

The reason for the segmentation pattern observed in relation to *attempt* might be that this element does not always exhibit an overt action and that it can be perceived through abstract concepts such as a plan that the protagonist has in their mind as a thought or intention. This was illustrated with extract 20, *Casino Royale,* in section 4.1, where multiple abstract candidates for the *attempt* element led to divergent segmentation decisions by the two annotators. However, the largest discrepancy for *attempt* (61.327 seconds) was observed in extract 2 *(Johnny English Reborn*), because Ann1 identified two sets of *attempts* (one for each protagonist) as a result of interpreting the story as two parallel narratives. An example of similarity between the annotators occurs in extract 5, from *The Guardian* (Figure 4.10). In this extract an overt action, i.e. a visual cue, can be identified as one protagonist's *attempt* to apologise, namely the moment when he anxiously looks down expressing regret, followed by scratching his forehead, which can be interpreted as a preparation for making an apology. The annotators' respective segment boundaries for *attempt* are very close, suggesting that the visual cue provided by the protagonist's body language has fostered similarity here.

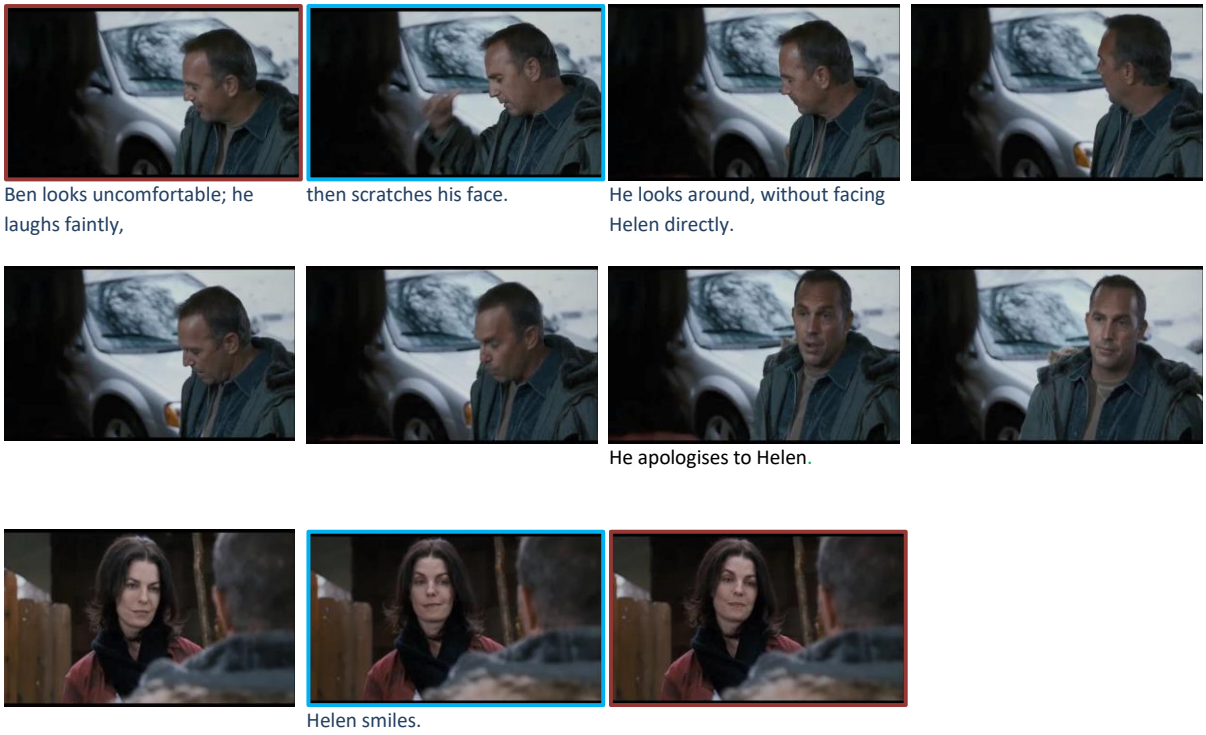

Figure 4.10 The Guardian_ 'Attempt' (Ann1 (Purple frames): time in: 00:00:22.749, time out: 00:00:35.370; Ann2 (Blue frames): time in: 00:00:25.568, time out: 00:00:34.924)

4.2.5 Consequence

The SG element *consequence* is defined as "an event, action, or endstate which marks the attainment or nonattainment of the protagonist's goal" (Stein & Nezworski, 1978:182). In our data, this SG element elicited the second highest discrepancy between the two annotators (M=6.795, SD=8.118).

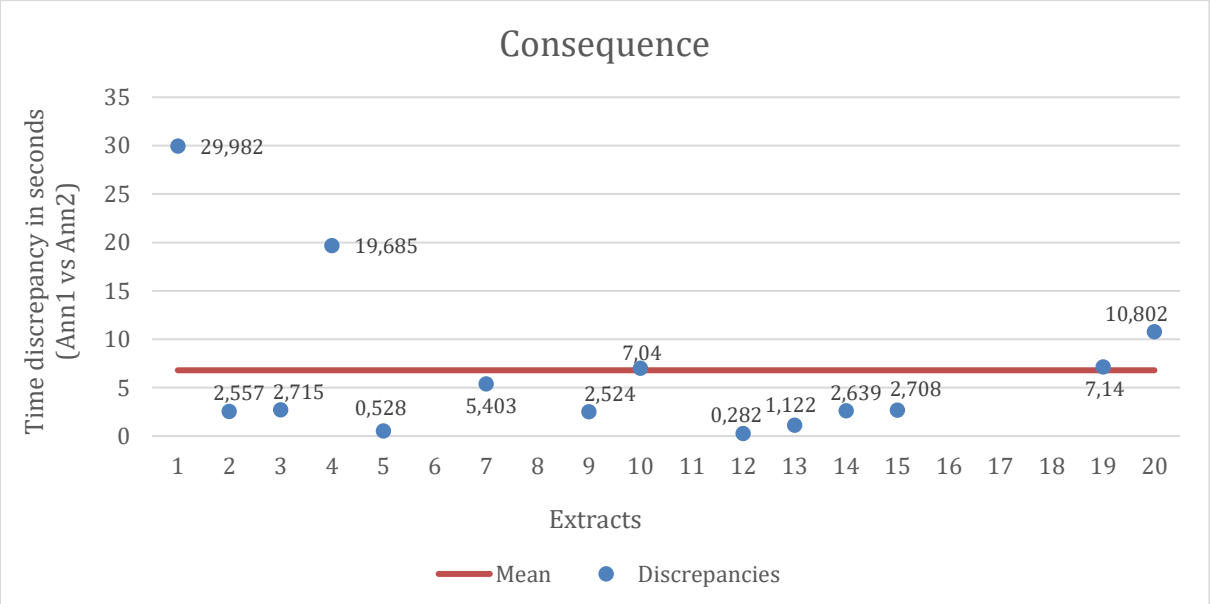

Figure 4.11 Duration time discrepancy between Ann1 & Ann2 for 'Consequence' (Mean: 6.795, Standard deviation: 8.118)

No	Extract title	Extract ID	Discrepancy	No	Extract title	Extract ID	Discrepancy
1	Extremely Loud…	[100900]	29.982	11	Goal	[201313]	N/A
2	Johnny English Reborn	[201500]	2.557	12	Memoirs of a Geisha	[101803]	0.282
3	The Aviator	[103312]	2.715	13	Click	[000703]	1.122
4	The Devil Wears Prada	[203409]	19.685	14	Bruce Almighty	[000501]	2.639
5	The Guardian	[203614]	0.528	15	Being Julia	[200310]	2.708
6	The Help	[203701]	N/A	16	The Aviator	[103311]	N/A
7	The Social Network	[004303]	5.403	17	Pretty Woman	[102404]	N/A
8	An Education	[000200]	N/A	18	The Matador	[004100]	N/A
9	Sex and the City	[102706]	2.524	19	The Forgotten	[003502]	7.14
10	The King's Speech	[004001]	7.04	20	Casino Royale	[200609]	10.802

Table 4.7 Extract IDs with the relating discrepancies for 'Consequence'

As can be seen in Figure 4.11 and Table 4.7 above, the annotators generated the most noticeable discrepancy in extract 1 (*Extremely Loud…*; 29.982 seconds). For Ann1, the fact that the woman has no answer to the boy's query represented the *consequence.* However, to Ann2, the woman declining the boy's request to kiss her was more narratively salient. The relating screenshots are provided below:

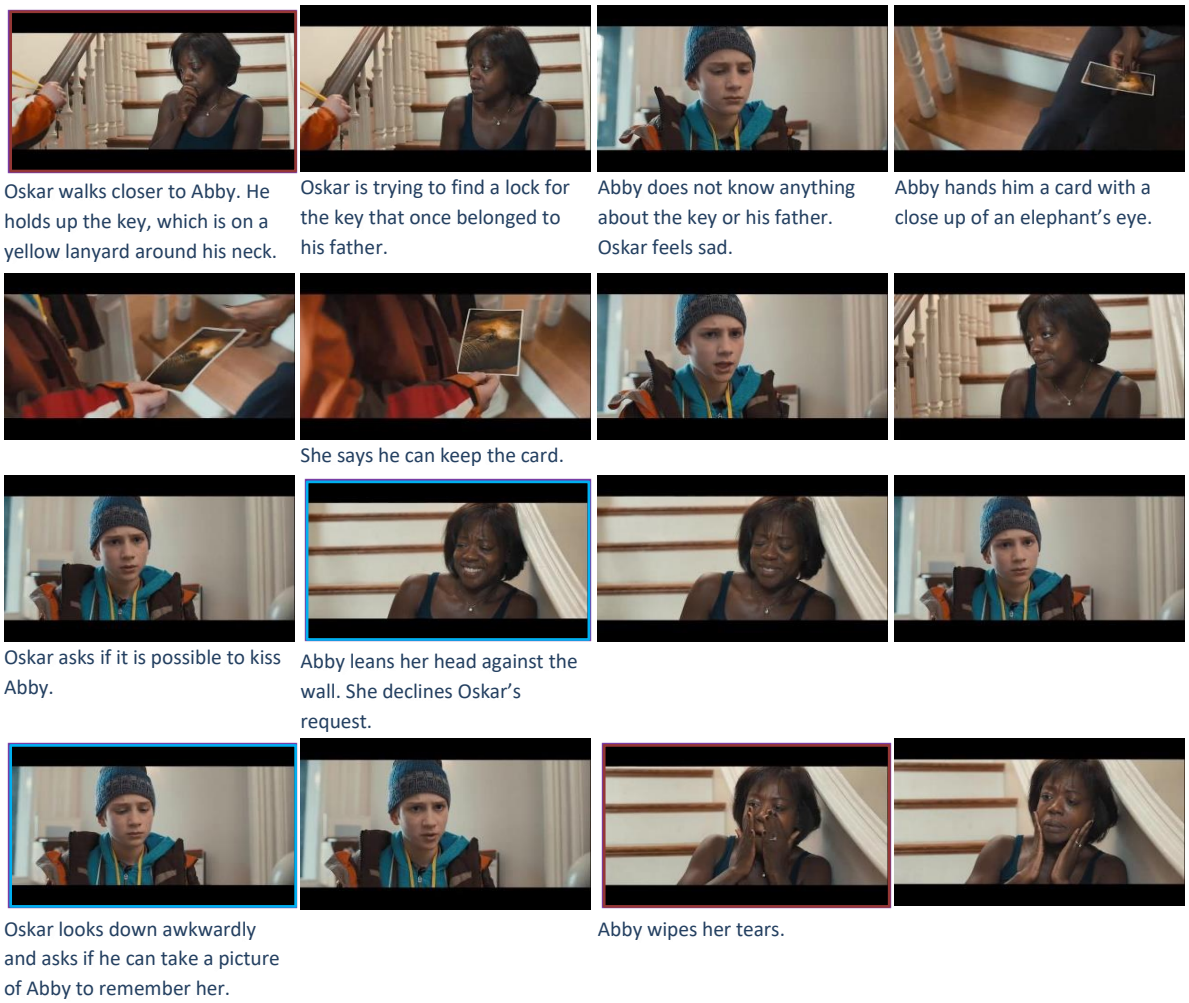

Figure 4.12: Extremely Loud…_ 'Consequence' (Ann1 (Purple frames): time in: 00:00:42.660, time out: 00:01:17.920; Ann2 (Blue frames): time in: 00:01:07.072, time out: 00:01:12.350)

4.2.6 Reaction

The final SG element is labelled *reaction* and is defined as capturing "a character's response to the consequence or broader consequences caused by the goal attainment" (Stein & Nezworski, 1978:178). Since *reaction* is traditionally the concluding SG element, a higher level of inter-annotator agreement is to be expected. This is borne out in our data: With an average discrepancy of M=5.332 seconds (SD=6.796), this element is among those with the lowest discrepancy. Only six extracts pertaining to this element fall above the mean, suggesting broad agreement with only a few outliers skewing the results.

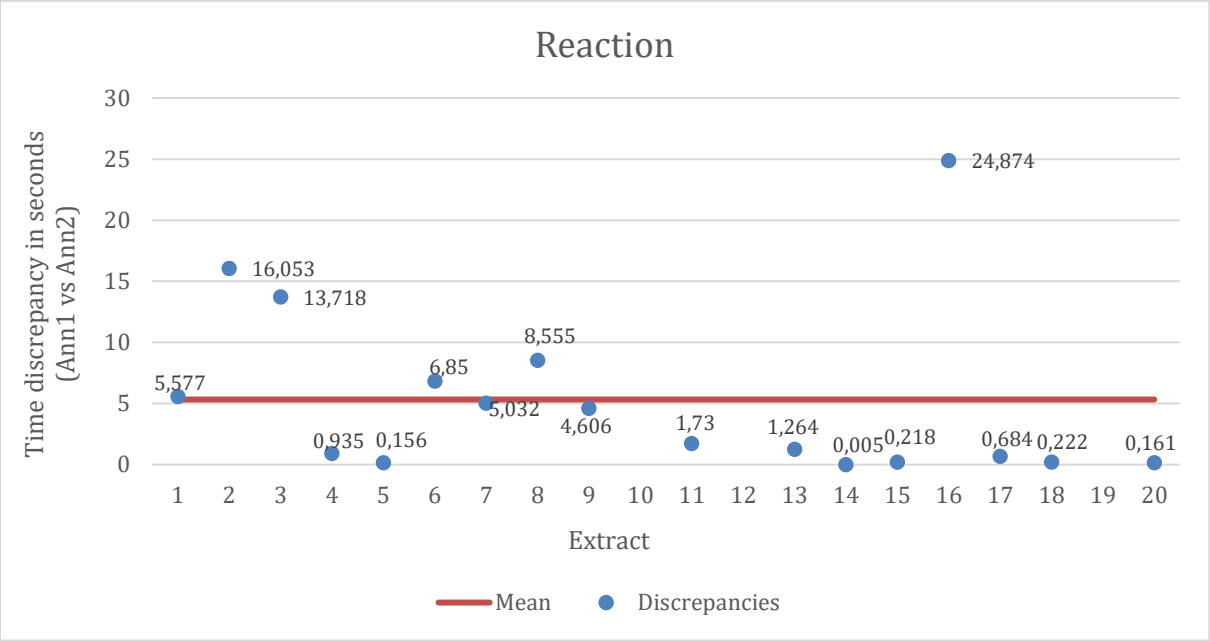

Figure 4.13 Duration time discrepancy between Ann1 & Ann2 for 'Reaction' (Mean: 5.332, Standard deviation: 6.796)

No	Extract title	Extract ID	Discrepancy	No	Extract title	Extract ID	Discrepancy
1	Extremely Loud…	[100900]	5.577	11	Goal	[201313]	1.73
2	Johnny English Reborn	[201500]	16.053	12	Memoirs of a Geisha	[101803]	N/A
3	The Aviator	[103312]	13.718	13	Click	[000703]	1.264
4	The Devil Wears Prada	[203409]	0.935	14	Bruce Almighty	[000501]	0.005
5	The Guardian	[203614]	0.156	15	Being Julia	[200310]	0.218
6	The Help	[203701]	6.85	16	The Aviator	[103311]	24.874
7	The Social Network	[004303]	5.032	17	Pretty Woman	[102404]	0.684
8	An Education	[000200]	8.555	18	The Matador	[004100]	0.222
9	Sex and the City	[102706]	4.606	19	The Forgotten	[003502]	N/A
10	The King's Speech	[004001]	N/A	20	Casino Royale	[200609]	0.161

Table 4.8 Extract IDs with the relating discrepancies for 'Reaction'

The highest discrepancy was observed in extract 16 (*The Aviator*), where the main protagonist is seen excessively washing his hands until they bleed. Ann1 perceived the *reaction* as starting when a sense of accomplishment is seen on the protagonist's face; by contrast, Ann2 regarded the *reaction* as beginning when the protagonist finished washing his hands and put the bar of soap in his pocket. In extract 12 *(The Memoirs of a Geisha),* timings for this segment were very close. However, the annotators' reasons for selecting the boundaries for this segment were not the same. To Ann1, it is the girl's expression of serenity and satisfaction when she makes an offering in the temple that marks the *reaction*. However, to Ann2, the

reaction was associated with filmic imagery, i.e. the girl seeing the Chairman again combined with the falling cherry blossoms to signify an end to childhood (Figure 4.14).

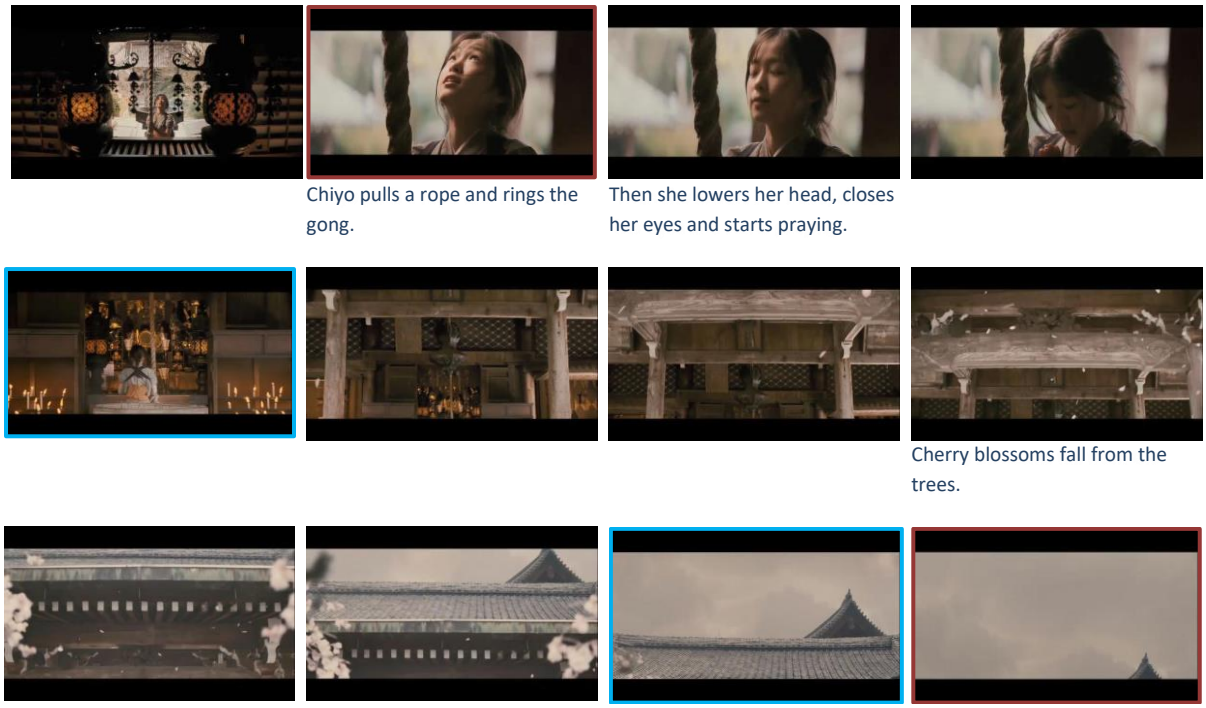

Figure 4.14 Memoirs of a Geisha_ 'Reaction' (Ann1 (Purple frames): time in: 00:00:42.758, time out: 00:01:01.691; Ann2 (Blue frames): time in: 00:00:51.320, time out: 00:01:01.858)

4.3 Segmentation: discussion

In this section, we discuss the findings presented above in more general, narrative terms, outlining the implications for our main questions, i.e. to what extent the discrepancies identified above are meaningful and what this means for the applicability of the SG approach in the context of training machines to understand and describe audiovisual narratives.

4.3.1 Broad segmentation agreement

Extracts with the greatest segmentation agreement between annotators are characterised by at least one of the following two features: (i) the same decisions in terms of the labels chosen and the order in which they have been applied; and (ii) similar 'time in' and 'time out' markers, suggesting only minor discrepancies in narrative cueing detection. In relation to inter-annotator time discrepancies, it has been acknowledged that annotator reaction speeds are likely to impact the selection of in- and out-frames for any given event. It is therefore unlikely that exact coincidence in timings will occur between two people even when identical cueing prompts are considered narratively salient. A margin of difference of one second or less is unlikely to correspond to a significant divergence of opinion in this regard. However, further investigations conducted to determine the split between audio and visual cueing prompts allowed us to explore this phenomenon further. This will be discussed further in section 4.4.

An example of broad segmentation agreement between annotators can be seen in an extract from *The Guardian* [203614] (Figure 4.15). Here, the SG labels were assigned by both annotators in the same sequence and with only minor time discrepancies:

Ann1			**Ann2**			**Discrepancy**
Time in	Time out	**SG label**	Time in	Time out	**SG label**	Duration
00:00:00.020	00:00:06.920	setting	00:00:00.000	00:00:07.588	setting	00:00:00.688
00:00:06.920	00:00:09.850	initiating event	00:00:07.607	00:00:10.274	initiating event	00:00:00.263
00:00:09.850	00:00:22.740	int. response	00:00:10.274	00:00:25.568	int. response	00:00:02.404
00:00:22.749	00:00:35.370	attempt	00:00:25.568	00:00:34.924	attempt	00:00:03.265
00:00:35.370	00:00:45.780	consequence	00:00:34.924	00:00:45.862	consequence	00:00:00.528
00:00:45.780	00:00:58.124	reaction	00:00:45.862	00:00:58.050	reaction	00:00:00.156

Figure 4.15 Broad Segmentation Agreement (Annotators 1 and 2), The Guardian [203614]

This example illustrates the utility of SG segmentation labels as a means of sectioning in the narrative into 'chunks' of data each of which denotes a milestone in the development of plot, and can be regarded as a cue for propelling the narrative towards its conclusion. Here, we can see that the discrepancy in duration of each of the labels is minimal, with 3.265 seconds being the greatest margin of difference (*attempt*) between annotators.

In more general terms, the analysis of segmentation duration timing discrepancies by segment label, shown in our scatter diagrams (Figure 4.4Figure 4.4, Figure 4.6, Figure 4.8, Figure 4.9, Figure 4.11, Figure 4.13), indicates that there are large clusters of results falling just below the mean, with fewer 'outliers' some way above the segment average. Thus, in the case of our two annotators it would appear that the average discrepancy in segment timings is skewed by above average performance of the few versus the lower discrepancies reflected in the many, suggesting that there is **more agreement than disagreement** in terms of segmentation allocation and labelling. The implication would therefore seem to be that there is merit in using SG to define the shape of narrative.

4.3.2 Minor segmentation discrepancies

In other cases, discrepancies between annotators took the form of minor differences in the attribution of segmentation labels. Rather than be restrained in the application of SG by the rigid application of standard schemata (Rumelhart 1975; Stein & Nezworski 1978; Mandler & Johnson 1977, 1980) which would have forced our annotators to 'shoehorn' the narrative to into an artificial and inflexible construct, they were allowed free rein to apply the Stein & Nezworski (1978) segmentation labels as they deemed most appropriate. Repetition, re-ordering and omission were all permitted as strategies to ensure the SG labels were applied to each narrative in the most meaningful manner. We considered this experimental approach to be necessary for evaluation purposes, since it was important to be open to the possibility that modifications might be necessary between a system that was originally developed for textual analysis, and its application to more dynamic multimodal material. The compromise this elicited was that some labelling sequences contained discrepancies between annotators. Conceptually, in many cases, the decision-making was similar but levels of granularity, for instance plot and sub-plot, or the subtle differences between *consequence* and *reaction,* were perceived differently and segmentation labels applied accordingly.

In our extract from *The Help* [203701], for example, both annotators assigned the same labels across the narrative: *setting/initiating event*, *internal response*, *plan/attempt* (see below), *consequence* and *reaction*). However, the latter two labels – *consequence* and *reaction* – occur in both possible ordering permutations, with Ann1 retaining the original SG order, and Ann2 reversing the order, albeit at close to identical timings (Ann1 'time in'= 00:00:16.575; Ann2 'time in'= 00:00:16.541). In this instance, *consequence* and *reaction* were not readily discernible one from the other. Indeed, this was observed throughout the annotation process, with the *consequence* of an action even being the *reaction* itself. Our recommendation would therefore be to consider a merging of these two segmentation labels when applying SG schema to moving imagery since, unlike written texts, the *consequence* does not always become evident before we witness the protagonist's *reaction.*

Ann1			**Ann2**		
Begin	End	**Label**	Begin	End	**Label**
00:00:00.010	00:00:06.198	setting	00:00:00.000	00:00:07.023	setting/init. event
00:00:06.198	00:00:10.214	init. event			
00:00:10.000	00:00:14.412	int. response	00:00:07.023	00:00:10.720	int. response
00:00:14.412	00:00:16.575	attempt/plan	00:00:10.720	00:00:14.559	plan
			00:00:14.559	00:00:16.541	attempt
00:00:16.575	00:00:24.027	reaction	00:00:16.541	00:00:20.505	consequence
00:00:24.027	00:00:34.940	consequence	00:00:20.505	00:00:28.056	reaction

Table 4.9 Example of variation in segmentation labelling: The Help [203701]

A further cause of segmentation discrepancy was found in the differences in interpretation of narrative on the part of the annotators. In extracts where symbolism may be present, in particular, there were sometimes fundamental differences in the analysis of key narrative milestones (see *Memoirs of a Geisha,* Figure 4.14 above). As discussed, both perspectives are valid since the girl's act of praying is certainly narratively significant to the evolving storyline, but the symbolism of the cherry blossom may also have been foremost in the director's mind at this point. If human annotators fail to reach agreement on issues of this type, it certainly raises difficult questions for training the computer in narrative comprehension.

4.3.3 Segmentation timing and labelling discrepancies

In other extracts, there were timing discrepancies between annotators. An example can be seen in our clip from *Extremely Loud and Incredibly Close* [100900]. Two observations stand out in relation to this material: firstly, that the durations for each segment differ significantly between annotators, even though the overall duration of narratively relevant information is almost identical (1 minute, 27 seconds); secondly, Ann2 chose to differentiate between the principal protagonist's *attempt* to undertake an action, and his *plan* regarding the activity. This additional component to standard SG was first observed by Rumelhart (1975), where a *plan* to perform a narratively relevant act precedes the *attempt* to do so. In our application of SG to the MeMAD500 film extracts, we observed that *plan* and *attempt* were often discernibly different acts, while on other occasions either one could be present without the

other. In order to capture this subtle distinction, both *plan* and *attempt* were eventually permitted within our segmentation schema.

Variable segmentation timings between annotators is an interesting phenomenon, as intuitively it might be supposed that an audience understands the point at which plans are actioned, and consequences evidenced, in a broadly similar chronology. Certainly some of the observed discrepancies were the result of physical reaction times in segmentation boundary setting between annotators. In other cases, the discrepancies were greater. While it became clear across our selected film extracts that *consequence* and *reaction* may occur either independently one without the other, or together but in either order, these elements of plot resolution were almost always present. Yet in the case of *Extremely Loud and Incredibly Close*, the duration of these two segments, which both annotators agreed occurred in the same order, were substantially different (Ann1=00:00:44.389; Ann2=00:00:19.984). This can be seen using the 'time in' data, which shows that Ann1 regarded the *consequence* phase as beginning at 00:00:42.660, while Ann2 saw this as commencing at 00:01:07.000. In short, Ann1's *consequence* was subsumed in Ann2's *initiating event* and *internal reaction* segments.

Ann1			Ann2		
Time in	**Time out**	**Label**	**Time in**	**Time out**	**Label**
00:00:00.000	00:00:04.640	setting	00:00:00.000	00:00:21.000	setting
00:00:04.640	00:00:07.420	initiating event	00:00:21.000	00:00:43.771	initiating event
00:00:07.420	00:00:23.090	internal response	00:00:43.789	00:00:46.859	internal response
			00:00:46.859	00:00:59.771	plan
00:00:23.090	00:00:43.000	attempt	00:00:59.774	00:01:07.072	attempt
00:00:42.660	00:01:17.920	consequence	00:01:07.072	00:01:12.350	consequence
00:01:17.927	00:01:27.056	reaction	00:01:12.350	00:01:27.056	reaction

Table 4.10 Timing and Labelling Discrepancies: Extremely Loud and Incredibly Close [100900]

These 'fuzzy boundaries' between segments, and the degree to which they are open to interpretation by the viewer, hint at the difficulties likely to be incurred when attempting to develop AI models and train computers to detect narrative milestones in a consistent way for the purposes of sequencing between frames and shots. Shot changes, which provide a strong visual dynamic although not necessarily evidence of a narrative shift, could impact segmentation choices and confound boundary choices. This theory was not tested during our study, but would be an interesting element of any future research, especially as computers are already relatively well trained in detecting shot changes. There may be potential to assign shot changes which define a scenic shift as initial segmentation breaks which subsequently require post-editing through human intervention. However, at this stage we have simply trialled SG methodology without considering current automatic segmentation techniques, suggesting refinements and alternative approaches to narrative segmentation schemata in order to accommodate highly complex film narrative.

4.3.4 Repetition and variations in segmentation labelling

A further observation in applying SG segmentation to film narrative is that the complexity sometimes lends itself to repetition of certain labels. This is most likely to occur where there are multiple minor deviations from the main plotline. In the context of what are already short narrative extracts, these digressions might be regarded as 'micro-narratives' which loop out

of the main plotline, follow a short circuitous route, and then rejoin the principal narrative. Our extract from *Sex and The City* [102706] illustrates this point. While Ann1 chooses to label one *initiating event*, followed by one *internal response*, Ann2 subdivides the action into a series of micro-initiating events and associated responses. Hence, for instance, when Carrie throws a look at her husband to see if he is studying her, Ann2 considers this as an *initiating event*, which is met with the *internal response* of annoyance when she sees that he is fixated on the television. Each action Carrie subsequently takes to engage her husband in dialogue, and the response to that action by either herself or her husband, constitute separate *initiating actions* and *responses.* That is not to say changes in conversational turns represent new segmentation boundaries in all cases, but that in specific instances, the plot may be moved along in this way. On the other hand, Ann1 has taken a 'macro' approach to the same material, designating all of the conversational turns as one *initiating event* except the final one, which she considers to be the *internal response.* It is possible to argue that both approaches are legitimate, since they are conceptually the same: an introduction to some event which propels the narrative towards a climactic event: the *plan/attempt*.

Ann1			Ann2		
Time in	**Time out**	**Label**	**Time in**	**Time out**	**Label**
00:00:03.133	00:00:15.099	setting	00:00:00.000	00:00:10.388	setting
00:00:15.099	00:00:24.540	initiating event	00:00:10.388	00:00:18.777	initiating event
00:00:24.540	00:00:41.297	internal response	00:00:18.777	00:00:24.321	internal response
x	x	x	00:00:24.321	00:00:38.000	initiating event
x	x	x	00:00:38.000	00:00:41.159	internal response
x	x	x	00:00:41.159	00:00:55.212	initiating event
x	x	x	00:00:55.212	00:01:03.190	internal response
00:00:41.297	00:01:17.533	attempt	00:01:03.190	00:01:17.327	plan/attempt
00:01:17.533	00:01:27.597	consequence	00:01:17.327	00:01:24.867	consequence
00:01:27.597	00:01:40.902	reaction	00:01:24.867	00:01:42.778	reaction

Table 4.11 Levels of Granularity in Segment Designation: Sex and The City [102706]

This clip exemplifies a further issue arising from SG when it is applied to film and television narrative. Whereas in book narrative we are often privy to the internal narrative of one of the main protagonists either through first-person or third-person narration, film convention dictates that this is often achieved through a series of shot changes. For instance, in our *Sex and The City* extract, a series of *initiating events* are matched by *internal responses* from both the person initiating that event and the respondent to whom it is directed. We may also witness a protagonist speaking via an 'over the shoulder shot', with the recipient of that utterance having their back to the camera; the camera then flips to an 'over the shoulder shot' from the perspective of the recipient, showing their reaction to the original speaker's remark. In this way, the *initiating event* and the *internal response* (or simply 'response', since the reaction is not always internalised), can be shared between two or more characters. Traditional SG does not address this point because it was conceptualised as a tool for textual analysis, suggesting that an adaptation would be necessary in the case of multimodal narratives.

4.4 Audiovisual cues

4.4.1 Using audio and visual cues as an indication in segmentation shifts

As mentioned earlier, we chose to establish whether the cues used to generate a segmentation shift were, in each case, 'audio' or 'visual' in nature. Although it was acknowledged that audiovisual cues are often presented simultaneously, nevertheless an attempt was made to spot the most **prominent** cue for segmentation boundary decisions. Every segment in all twenty extracts was therefore investigated in detail to identify the prominent cue selected by both annotators. In general terms, both annotators tended to rely on visual cues more frequently than audio prompts (Table 4.12 and Figure 4.17).

Our three film extract categories—dialogue, monologue and nil speech—were anticipated to have an impact on audiovisual choices. In the case of 'nil speech', we expected that audio cues would have little impact on boundary demarcation choices. In practice, this was not the case: Ann2 used sound effects to cue certain segmentation choices, while Ann1 relied more on the visual cues. It is entirely possible that when a director creates a scene with no dialogue, even where there are sound effects, they ensure visual cues alone are enough to carry the narrative. As a case in point, in *The Matador* extract (nil speech), Ann1 used only visual cues throughout the segmentation process whereas, Ann2 used 40% visual and 60% audio cues.

For audiovisual productions, it can be argued that human beings generally consider both audio and visual cues to be narratively salient when making decisions about segmentation boundaries. Taking *The Aviator* [103311] as an example, the consequence of one character excessively washing his hands until they bleed is portrayed through both visual and audio cues almost simultaneously (the facial expression that shows discomfort and the sound the character makes out of pain; Figure 4.16). Nevertheless, one annotator still found an audio cue more persuasive as a segmentation cue than the visual prompt.

Audio: Rubbing hands with soap. | Audio: Water running. | Audio: Fast hand rubbing. | Audio: Manic hand rubbing.

Audio: Obsessive rubbing sound stops. | Audio: [Catching breath]. | Audio: "Ahh!" [Pained expression]. | Audio: Silence.

Figure 4.16 Screenshots and audio summary of The Aviator [103311] extract relating to visual cues

4.4.2 Cueing prompt analysis: Audio vs. visual cues as segmentation markers

In the course of the discussions between the annotators regarding the allocation of SG segment labels, some consideration was given to the effect of film dialogue on the selection of segment boundaries. Both annotators acknowledged there was an initial attraction to the pragmatic boundaries represented by conversational turn-taking in the film material, as markers for segmentation breaks; there was also an awareness of the need to avoid being distracted by the audio descriptions which would have represented a confound. Nevertheless, AD was immediately discounted as a distraction from making diegetically informed narrative segmentation choices due to its asynchronous nature. Dialogue markers, on the other hand, became a topic of particular interest and their impact on the choice of segmentation boundaries debated at some length between annotators. Both annotators noted it was likely that some segmentation choices were affected by conversational turns within the film dialogue and that these could have been dominating segmentations choices.

Extract	Ann1				Ann2			
	Vis.	Aud.	Visual %	Audio %	Vis.	Aud.	Visual %	Audio %
Extremely Loud…	2	4	33	66	2	5	29	71
Johnny English Reborn	6	3	67	33	4	1	80	20
The Aviator-13	5	1	83	17	3	3	50	50
The Devil Wears Prada	5	1	83	17	7	2	78	22
The Guardian	4	2	67	33	3	3	50	50
The Help	3	3	50	50	3	3	50	50
The Social Network	4	2	67	33	3	3	50	50
An Education	3	2	60	40	4	2	67	33
Sex and the City	5	1	83	17	6	4	60	40
The King's Speech	2	3	40	60	1	5	17	83
Goal	4	2	67	33	4	3	57	43
Memoirs of a Geisha	3	0	100	0	4	1	80	20
Click	3	3	50	50	0	6	0	100
Bruce Almighty	3	2	60	40	3	3	50	50
Being Julia	5	1	83	17	5	1	83	17
The Aviator-12	10	0	100	0	6	1	86	14
Pretty Woman	5	0	100	0	6	0	100	0
The Matador	4	0	100	0	2	3	40	60
The Forgotten	4	1	80	20	5	0	100	0
Casino Royale	6	0	100	0	5	1	83	17
Totals	86	31	**74%**	**26%**	76	50	**60%**	**40%**

Table 4.12 Division of SG Segmentation Cues by Annotator, Audio vs. Visual

Dialogue	Monologue	Nil Speech

Ann1 *Ann2*

Figure 4.17 Audio-Visual Segmentation Cueing Split, Ann1 vs. 2

However, counterintuitively, considering the dominance of dialogue in film plotting, there was a clear bias towards visual cueing for the determination of segment shifts in the case of both annotators (Ann1: 74% visual, 26% audio; Ann2: 60% visual, 40% audio). Given earlier discussions, this result was somewhat surprising. The natural conclusion would seem to be that visual elements like body language, facial expression and narrative actions carry a greater weight in the development and transitional determination of plot than either annotator had anticipated. Naturally, we cannot suggest this finding applies universally to audiences, and it is entirely feasible that our results could have been skewed by the type of film material selected, or indeed our annotators' own personal visual bias. However, the source of segmentation cueing is certainly an area of investigation worthy of further attention. There may also be some correlation between an individual's preferred learning style, with both of our annotators acknowledging that they are visual learners.

One further point of note is that audiovisual material generally combines simultaneous audio and visual cueing to create meaning, and so establishing which of these channels is more dominant for the purpose of developing narrative is not always a simple matter. Nevertheless, if it were possible to prove that human beings place a bias on visual information over verbal when decoding narrative, this information could be useful for developing future machine-based models. The implication would be that any element of mathematical weighting introduced between computer vision based calculations and those determined on the basis of automatic speech recognition and topic detection, should favour the former.

4.5 Summary: Application of Story Grammar to automating narrative segmentation

Our investigations into the application of SG for modelling human narrative sequencing have shown that methods originally proposed as a way of capturing textual plotlines are not suitable for direct transfer to multimodal material without adaptation. Segmentation rules will differ for moving imagery where markers signifying a narrative shift in storytelling may come from either audio (dialogue, sound effects, musical scoring) or visual (actions, body language, facial expression, text on screen) sources. Moreover, audiovisual narrative tends to be complex in nature, with plots and sub-plots often running simultaneously, making a sophisticated modality of storytelling and narrative exposition. Yet, even in the most complex of scenarios, there is generally an underlying thread that can be captured using the SG schema.

Studies using SG for assisting recall concluded that children follow a natural SG in re-telling stories regardless of the order in which they were originally relayed (Hayes and Kelly, 1985). Moreover, the same study suggested that both adults and children displayed better narrative recall for *setting* and *outcome* segments than they did for *endings* or *reactions* (1985:346). This pattern was partially replicated in our study where both annotators found the *setting* sequences showed minimal temporal discrepancy; however, we also found low discrepancy in the *reaction* category.

Our finding that visual cueing tends to dominate segmentation decision-making is also curious in that it re-affirms the findings of one study which reported that children retain visual information more readily than auditory information (Hayes & Birnbaum, 1980), and that those parts of narrative which are most visually defined are *setting, attempt,* and *outcome* ("actions, consequences and the background in which they occur", Hayes and Kelly, 1985: 347). Again, these are the segments which invoked greatest agreement between annotators. By contrast, it was noted that "reactions and endings are often dependent on verbal discourse for presentation" with dialogue enhancing inferential reasoning (1985: 347).

Our conclusion is therefore that SG is a useful means of expressing the development of plot across a narrative arc, whether at the feature film level or at the level of micro-narratives as illustrated by this study. The emphasis on visual storytelling, and the degree of acquiescence on the *setting* and *reaction (*outcomes) labels, would suggest that machine-based learning models might be best focused on determining these two categories of narrative segmentation first, with the addition of ASR and topic detection at a later phase to fill in the gaps in less readily accessible segments (*consequence, internal responses*).

Next Steps

As discussed previously, automatic segmentation by the machine currently takes place at the shot level (see Deliverable 5.4), with each shot change heralding the beginning of a new segment. Improving results for face and object detection should enhance the sensitivity of segmentation with, for example, a change of principal protagonist prompting the start of a new segmentation. There are two potential routes to applying the findings of this study to future machine models, as below.

Method 1: Three-Part Narrative Segmentation
Perhaps the best approach for future machine learning might be to model for determining the *setting* and *outcome* segments, and rendering everything that falls between them *attempt*. In effect, this would be following Rumelhart's (1975) *event-goal-outcome* story grammar, which is essentially a simplified version of Stein and Nezworski's (1978) schema, with a more fluid core and less rigid boundaries. The very high level of agreement between our two annotators in the placement of Rumelhart's segment boundaries reinforces the value in testing this approach.

Method 2: Deeper Dive into Narrative Segmentation
However, segmenting for narrative will ultimately require a more elaborate approach, with scene detection and dialogue also playing a significant role. The segment shift from *setting* to *initiating event* could be detected visually in many cases, since providing an 'establishing shot' is a common directorial device to locate action in a particular time (e.g. darkness signifies 'that night' or 'later that day') and place ('in the park').

In our example from *The Devil Wears Prada* the *setting* includes a brief shot of the main character on her way home in a taxi, followed by an establishing shot of the inside of a small apartment in which a young man sits alone on a sofa. A woman dressed in party clothes walks into the room carrying a cupcake lit with a single candle and says “Hey”, followed by “Happy Birthday” to the man who appears depressed. The narrative action then commences with a short dialogue between the two characters. From this *setting* it is possible to establish the location (small apartment), the narrative moment (it is his birthday, hence the cake with candle) and the sub-narrative (she has been out alone on his birthday). The *initiating event* takes place when the woman apologises to the man. From this scene, the machine might therefore be trained to treat the taxi and brief interior apartment shots as the *setting*, and to shift to *initiating event* once the dialogue begins. In a similar vein, it is likely that the final shift from *plan/attempt* to *consequence/reaction* would be better determined in the first instance by training the machine to look for a conclusion to the dialogue, since visual elements alone may not change sufficiently to predict this narrative shift.

PART B

5 Guidelines

5.1 Introduction

Intended for broadcasters, audiovisual archivists, video/audiovisual content and platform providers, and developers, the present guidelines consider approaches to automating the description of video content (moving images) and audiovisual content (moving images combined with other modes of expression) for different purposes. One of the objectives is to facilitate media access for audiences with additional accessibility needs, especially sight-impaired audiences (audio description). Relying heavily on human resource, audio description is currently an expensive part of the post-production process for media companies, making it challenging to provide comprehensive media access in line with legal requirements. The recent increase in user-generated audiovisual content has created a further challenge for media access. The other purpose is the description of visual and audiovisual content for archive retrieval in the broadcasting context, to facilitate re-use of content internally or for re-sale to other media companies. Both types of description create similar questions about available resources, making automated methods an appealing proposition.

Research on automating the description of video scenes (**automated metadata extraction**) has intensified and has begun to show moderate success. The question to what extent automated methods can be drawn upon to produce descriptions for the above purposes without compromising quality and user experience is emerging as an economically and socially important question for research and practice.

An important step is to acknowledge the contribution to be derived from an analysis of human descriptions of video/audiovisual content, which has the potential to propel automated metadata extraction beyond standard object-and-action recognition tasks into the realm of multi-character, sequentially relayed narrative. In line with this, the primary focus of this guide is to outline: (a) how human approaches to constructing and understanding narrative are currently reflected in (semi-)automated approaches to describing audiovisual content and (b) how human approaches of multimodal meaning-making can be used for improving the automation of video captioning in the fields of media archive retrieval and media access.

Although the different purposes of description overlap to some extent, the main driver for description of archive material, i.e. the re-use of the content, tends to make these descriptions more “literal” or factual than audio description for sight-impaired audiences, which is often “narrative” or figurative. As the (semi-)automation of archive material description is therefore likely to be a more achievable goal in the shorter term than a model for generating elaborate audio description, these guidelines are focused on the generation of simple, descriptive video captions of a type most suitable for use in the context of archive content tagging and description. They do, however, also represent a first step on the long road to developing (semi-)automated methods for describing audiovisual content for

audiences with additional accessibility needs, whether physical (sight-impaired) or cognitive (learning difficulties, language-related disabilities, atypical cognitive frameworks). Since audience-oriented descriptions require a far more sophisticated type of cognitive processing of the source material than content descriptions designed for content retrieval, these guidelines should be expanded in the future, in parallel with the growing sophistication of machine outputs.

The guidelines have been informed by research into human approaches to understanding and constructing video descriptions, analysis of emerging automated video captioning, existing guidelines for human-derived CD, as well as national and supra-national guidelines and standards for human AD (e.g. OfCom, AENOR, ISO).

5.2 Computer Modelling Human Understanding and Constructing Video Descriptions

This section gives an overview of the human approach to understanding and constructing narrative/descriptions, outlines how their production can currently be supported by automation, how this is currently used and how (semi-)automated approaches can be implemented. It does so for different levels of human understanding, to highlight the extent to which these can be implemented, and the obstacles that need to be overcome.

5.2.1 Level 1: Key Elements

The human approach:
For computer-generated video captions to become more 'human-like' it is necessary to acknowledge that human meaning-making occurs at many levels and on many planes of comprehension and inference. As humans, we begin by establishing the basic facts to which we have assigned the acronym 'CALMO': who are the main protagonists (*Characters*); what are they doing that is narratively relevant and suggests the direction in which the narrative might evolve (*Actions*); where is this action taking place (*Location*); how might we interpret the emotional temperature of the piece (*Mood*); what props occur in the scene which might be considered narratively significant (*Objects*)? At the most fundamental level, these are the questions which allow us to engage with the unfolding plot and infer further salient facts such as the way characters are connected or related, any underlying themes or motifs, a sense of the narrative poignancy of the piece, and so forth. These five 'key elements' can be considered the narrative building blocks from which plot and sub-plot are determined, and without which traditional storytelling cannot exist.

How can this level be supported by automation?
The human process of identifying these 'key elements' can be simulated as a simple **metadata or 'tagging' operation** (*character, action, location, objects, mood*) performed on audiovisual material, drawing on current approaches to automatic topic segmentation (i.e. topic change detection) and topic modelling. Much less onerous than the process of modelling for automatic video captions, automated tagging can serve as a basic level of description. It can

be further improved through human intervention, which could be carried out alongside other post-production operations such as subtitling.

How can automation at this level be used?
In the context of **archive retrieval**, this specific form of metadata could be the first step in filtering/identifying film material at the start of a data 'search' function, and could be a 'quick win' for human annotators to find **less valuable assets**. This differs from the metadata generated at present and imported into the Flow platform direct from the broadcaster's media files (Avid). The level of human intervention when using our suggested 'key elements' can be adjusted according to the purpose and importance of the material. A considerable advantage of this approach, which combines automation with human input, is that, once a sufficient volume of material has been tagged in this way, e.g. by human archive journalists, this material can serve as **training data** for further machine learning.

How can these solutions be implemented?
The MEMAD approach to implementing auto-generated tags (in the Flow tool) is to offer the human operator automatically generated keywords, which can be accepted, deleted, edited etc. Project partners Limecraft have implemented this approach in the Flow platform, with archive workers being offered possible character, or even actual, names. Historically, this worked for 'named entities' only, such as famous politicians or celebrities, but now includes 'non-persons' (i.e. individuals for whom there are not vast training datasets of images available on the internet). These faces can be tagged as recurring characters and a name later inserted into the Flow tool by the archive journalist; the tool then designates that name to all occurrences of the individual across the extract. Our partners at EURECOM are responsible for developing face recognition models which introduced this type of 'non-person' identification to the prototype.

5.2.2 Level 2: Content Descriptions

The human approach:
Human comprehension is continuous and cumulative. It moves seamlessly from establishing 'key elements' of an unfolding narrative to gaining an appreciation of the sequence of events by constructing a sense of the narrative action as it occurs across time and space. At this stage, engagement with the storyline is at a basic 'what is happening' and 'in what order' phase. This can be captured in 'content descriptions' in which actions and characters, locations, objects and the mood of the piece are detailed in the order in which they occur, described in a non-interpretive manner.

How can this level be supported by automation?
Although at present many of the simpler tasks involved in automated video captioning, e.g. object recognition, character identification and action detection, remain unreliable, the stage of human information processing and meaning-making outlined above can to some extent be

achieved by current computer vision algorithms (such as the *DeepCaption* tool developed by our partners at Aalto University).

One of the main difficulties with producing machine-generated captions is that the models and algorithms used in their production draw upon large-scale training data in order to learn the required behaviours, primarily, object recognition and feature extraction. While the available datasets (e.g. MS COCO, TGIF, Visual Genome) are scaled appropriately, the crowdsourced nature of the captions and the banks from which images are drawn often results in topic bias and inaccuracies of description. Further confounding matters, many of these datasets describe still images or very limited moving image sequences containing only simple actions. None of them offers data of a level of sophistication approximating the moving images found, for example, in film and television presentations.

Where can automation at this level be used?

However, in spite of its current shortcomings, automatic video captioning can be considered for **archival purposes** for retrieval of AV content based on text search, as long as risks arising from erroneous descriptions are and/or can be mitigated, especially through human post-editing of auto-generated captions. The MeMAD work has also shown that the accuracy of auto-generated captions based on computer vision/object recognition algorithms can be improved by complementing these algorithms with information derived from facial recognition, speaker diarisation/vocal profiling, Automatic Speech Recognition (ASR) etc. Our team at Lingsoft has experienced good results with speaker diarisation, although the accuracy varies depending on the language spoken (Finnish being more successful than Swedish, for example). Work on vocal gender profiling has also advanced, with INA colleagues making great strides in performing male/female audio segmentation during political debates (Doukhan et al. 2018). To what extent this type of description/captioning is currently useful for **improving media access** for audiences with additional accessibility needs is debatable. This will be discussed further under Level 5 below.

How can these solutions be implemented?

In relation to archival purposes, tool development in this area is fast and highly specialised, meaning that a modular approach may currently provide the most flexibility. Rather than building 'super-algorithms' integrating all of the specialised areas, with a potentially high error rate, a modular approach that supports human intervention in each area is more transparent and fruitful. This is the approach which has been implemented in the MeMAD Flow tool and which has been informed during the developmental phases by frequent input from the lexical and narratological research undertaken at Surrey. Future plans involve a significant move towards integrating narrative segmentation into Flow, so that shifts in storytelling are retrieved and labelled for human operators to access and possibly edit (see Deliverable 5.4).

5.2.3 Level 3: Cohesive Ties and Establishing Relevance

The human approach:

The consumption of narrative storytelling is an act of decoding with clues to be gathered and assembled in order to reach an understanding of the storyteller's intent. The human mind takes the events it witnesses at Level 2 (above) and searches for verbal and visual cues which link characters and actions across time and space (cohesive ties), applying powers of inference and relevance-seeking to make sense of unfolding events and determining narrative saliency. In human descriptions, these actions would equate to an 'event narration' stream: an annotators' interpretation of the cues and prompts found in the source material which are subsequently used to make sense of the wider plot by reference to ongoing and past narrative events. An example of this is the falling cherry blossom in the extract from Memoirs of a Geisha (4.3.2), which requires a non-literal interpretation to gain access to the subtle narrative shift indicating an end to the main character's childhood, and linking to her adult future as a geisha.

How can this level be supported by automation?

In computer modelling terms this phase aligns with efforts to plot the actions of protagonists (including gender detection of the type undertaken by vocal means at INA, and face recognition to which EURECOM, INA and Aalto have all made significant input) and objects across frames, shot changes and scenes, using computer vision techniques. It also fits with an enrichment of Level 1 'metadata', which can add more detail to the identification of recurring people, objects and locations, in partaicular. However, this is only a first step in establishing continuity across the piece. As a step towards representing plot as occurring within the context of a temporal continuum (see Level 4 below), a variety of referential expressions including pronominalisation cues are used by human editors in both written and spoken narrative to avoid unnecessary repetition. Computer-generated descriptions, by contrast, fail to apply personal pronouns in a meaningful way; indeed, pronouns are only used rarely, and then within the context of a single video frame caption. Across-frame captioning is outside the computer's capacity, hence nominal and pronominal cohesion do not occur at the scene-wide level, with the result that many of the simpler clues to continuity of action are absent. More reliable facial recognition, and the attribution of identifiers to 'non-persons' (see Deliverable 3.3, forthcoming) will be a welcome first step in the direction of assigning pronouns across machine-produced video description captions. It may then be necessary for human operators to select ('M'/'F'/Plural) type labels and apply them to the persons identified, in order that relevant pronouns are consistently applied across the whole narrative by the machine. This is something that could easily be built into future iterations of the Flow platform.

Where can automation at this level be used?

Despite initial promising steps towards automating audiovisual storytelling on a more comprehensive scale than by identifying key elements (level 1) and/or deriving basic descriptions of individual events (level 2), automatically generated video captions currently

fail to achieve human-like accuracy, cohesion and relevance. However, work conducted in the MeMAD project has demonstrated that accuracy and cohesion can be improved when computer vision models and algorithms focused on object recognition are complemented by facial recognition, speaker diarisation, vocal profiling, ASR etc. Selective approaches, aimed at distinguishing salient and relevant information, are, however, likely to be beyond reach in the field of automated metadata extraction for the foreseeable future.

How can this level be implemented?

Achieving level 3 will require human supervision and intervention in the description workflow, as human-machine interaction seems best suited to achieve the balance between the growing time pressure in the media / broadcast sector and the ensuing need to accelerate workflows with quality and accuracy. The modular approach pursued in the MeMAD Flow platform (Deliverable 5.4), which makes outputs of different workstreams available to a human editor is therefore deemed to be a more fruitful approach in the short term than higher levels of automation. We aim to evaluate the utility of the Flow platform in this functionality with archive journalists at Finnish broadcaster YLE in the final months of the project. If this study proves successful, investment in the training of **human (post-)editors** who are able to process automatically generated output, as an example of fitting the platform tool into the real-life commercial scenario, effectively will be an important area for short-term development.

5.2.4 Level 4: Creating a Narrative Framework

The human approach:

As the highest level of human inference and meaning-making, the cues and cohesive ties identified between characters, actions and objects in Level 3 are collated and sorted to determine the development of narrative. The general shape of a plotline is first sought (event, goal, outcome), followed by the dramatic milestones (settings, events, reactions, attempts, consequences, outcomes) which are assembled in the human mind to establish narrative form and progression.

In the early stages of the MeMAD project, YLE production staff were observed engaging in the process of narrative storytelling in the edit-suite (Figure 5.1).There were many parallels between this human activity and the type of Level 4 meaning-making described here. At YLE, editors organised their documentary film rushes by naming and numbering each segment of the narrative at the 'rough cut' phase of editing (i.e. where the raw film footage must be placed into a meaningful order). As a way of systemising the kind of segmentation of narrative milestones that occurs effortlessly in the human mind, and is reflected in the frameworks supplied by SG, YLE editors named and numbered each phase of the narrative on a 'post-it' note, and displayed these on the edit-suite wall (
Figure 5.2). When it was decided that the order of one or more segments should be re-arranged to produce a more coherent or naturally flowing narrative, the numbered 'post it' note was retrieved and resituated on the wall to reflect the 'new' segmentation order. This process was repeated many times until the editor was content that the story milestones and

narrative arc were optimised to tell the story she was editing in the desired manner. This post-it note 'paper rough cut' (terminology commonly used in the film editing industry to signify an edit made on paper before transferring to the screen) was then replicated on the Avid film editing system (Figure 5.3).

Figure 5.1: YLE Editing Process

Figure 5.2: Post It Notes Showing Segmentation ('Paper Rough Cut')

Figure 5.3: Corresponding Film Segments Created on Screen (circled)

How can this level be supported by automation?

The ultimate goal in the drive to automate video captioning – and, indeed, captioning of multimodally constructed audiovisual content – remains to achieve narratively coherent, contextually sensitive and fully sequenced computer-generated storytelling. In order to reach this point, continuity of character identification and naming of the type trialled by EURECOM, would need to be established between shot- and scene-changes, taking into account variations of camera angle, cinematographic staging and mise-en-scène, and overcoming confounds such as changing appearances (e.g. differences in costume, hair styles, body profile etc.). Object tracking, and an understanding of the relationship between multiple objects, or objects and characters, would be essential to the provision of continuity and the development of sequenced storytelling. Our work analysing the machine descriptions from the computer vision team at Aalto has shown that this remains a major challenge and a resolution is largely

dependent on bigger image datasets becoming available in the future. Moreover, temporal sequencing takes on particular relevance both in terms of denoting the chronology of plot and in defining the general shape of the story, or how it is told. While temporal words currently occur in the vernacular of computer-generated captions, they have a spurious relationship with the types of temporal words human beings draw upon when conveying narrative coherence. As pointed out above, most available training datasets include still images or only limited moving images and can therefore not support the identification of markers of narrative connectivity.

Where can automation at this level be used?

The use of automated video descriptions at this level of sophistication would be useful in archival and media access contexts but is currently beyond the reach of computer vision models.

How can this level be implemented?

Computer sequencing using face and object tracking, together with segmentation techniques which draw on the MeMAD findings using SG principles should progress this work. The MeMAD 'Flow' platform currently segments material via shot detection, but it is anticipated that further work on combining these methods with visual, and perhaps ASR-detected audio, narrative shift detection could produce tentative steps toward first narratively based segment creation. Separating out the establishing shots which comprise the *setting* phase from what follows, being one of the segments generating most agreement between our two annotators, is the suggested first goal (see D5.3, part 1). Furthermore, topic detection and clustering methods (EURECOM) have been debated with a view to establishing major changes in narrative direction within moving images; while computer vision techniques developed at Aalto, based on the grouping together of image frames which appear very similar in order to detect a sizeable shift in image type portrayed in subsequent frames, have also been discussed as a way to detect narratively-driven segmentation boundaries (scene classification and shot merging). These latter two developments have been under discussion between Surrey, Aalto and EURECOM for some time, and are now starting to be realised. It is hoped that Surrey's work over the final months of the project will produce further insights from an extended SG study, which can be fed back into the EURECOM /Aalto segmentation activities.

6.2.5 Level 5: Audio Description

Audio description (AD) is a type of video description that sits alongside the narrative pathway outlined above. It relies not only on a thorough comprehension of plot but also on the extraction, omission, simplification, prioritisation and non-duplication of information to fit neatly within short gaps in the existing dialogue. Well-framed audio description should address the question: what do sight-impaired audiences need to know to make sense of this material that cannot be gleaned from the soundtrack?

This is a question that cannot be readily answered by neural networks in their current state of development. Computer-generated video captions do not currently meet legal requirements for media access (viz. meaningful description) and can therefore not replace human audio description as a service for sight-impaired audiences.

Semi-automation of AD is a more realistic goal, especially when content descriptions (see Level 2 above) are becoming sufficiently accurate and cohesive (i.e. are progressing to what we have labelled Level 3 above), as there may then be merit in passing them to human describers for **post-editing**. Automated captioning with post-editing workflows afford opportunities to enhance the machine's best efforts, with human-in-the-loop approaches not only being used to improve computer vision models but also to determine how human and machine intelligence can most productively and efficiently come together.

Combining human and machine endeavours in this way will also demonstrate to the human creators of AD that their involvement in developing or improving automated workflows will not mean that they are writing themselves out of their jobs, but that it will contribute to the development of automated methods for situations where professional AD is not available. 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.

At the same time, if automated approaches to the production of AD are pursued, the likely problems with the accuracy of automated video captions means that **risk mitigation** strategies need to be developed in relation to critical content (e.g. public health information), which should be identified and marked as unsuitable for automatic description.

Taking the path outlined here 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.

5.3 Key Areas for Improvement in Computer Modelling and Video Caption Automation

As a task list for moving computer description of narrative Level 1 to Level 4, and to make progress towards attaining Level 5, the following improvements to machine description models should be actively considered:

5.3.1 Efficient character identification and tracking

At the most fundamental level, automated video captioning and audiovisual storytelling relies on the correct identification of narratively significant protagonists. Gender labelling may not be wholly desirable in a political context but in archival retrieval and audio description scenarios the designation of male and female helps to quickly disambiguate between multiple characters. The choice of 'a woman' or 'a man' rather than 'a person' also sets the stage for future references to be pronominalised and therefore less repetitious (see gender

identification, above). In television archives, being able to distinguish between two or more characters by any distinguishing factor, whether gender, hair colour, eye colour, clothing style or any other feature, speeds up retrieval. This is important in commercial applications where time is money.

At the present time, certain factors appear to dictate gender labelling in an unreliable way: long hair suggests 'woman', short hair suggests 'man'; clothing such as trousers and jackets suggest 'man', and so forth. This needs to be rectified. As human beings we seldom misinterpret the gender of another person, even where contra-indications might be expected to confound our analysis. We can generally detect when a man is dressed as a woman, or vice versa, and distinguish that scenario from a man in a kilt or a woman wearing military fatigues. The computer is not yet capable of such subtle distinctions. Facial recognition, including facial contour profiling, and voice analysis, may assist with this task, as will named-entity recognition. If a character cannot be 'recognised' and named by reference to a sufficient depth of training data, then gender allocation alone will still assist with intra-narrative (local and global) cohesion.

Character identification	Improve gender and character identification via face recognition and voice diarisation methods.

5.3.2 Intelligent object recognition

Although object recognition is slowly improving in the context of still images, it continues to be a problem in general purpose (non-bespoke) computer-generated video captioning where objects are frequently mis-labelled. Atypical shot angles or anomalies of scale often act as confounds. There seems to be a lack of training data to rectify this at present, since many millions of images are required for an object to be consistently identified with precision given the potential permutations of angle, size, colour, form and so on.

Human beings determine the nature of objects both iconically, and by reference to context. We can all recognise our national post boxes by reference to their shape, colour, name of the postal organisation printed on the front, but when we are confronted by a letterbox of a different country, particularly if we cannot read the language, it may be confused for another object or overlooked entirely. However, if that unfamiliar letterbox is outside a post office, or we see someone posting a letter there, or the postman making a collection, the context tells us that this is likely to be a post box and not a litter bin. To some extent, contextualised object recognition may therefore compensate for the life experience a computer clearly lacks. Topic detection, which can be achieved either visually via scene analysis, or through ASR and dialogue 'comprehension', should produce computer models that are closer to human cognition. Objects detected by the machine could then be evaluated against the likelihood of their relevance to any given scene or topic.

Object Recognition	Enhance contextualised object recognition for greater sensitivity to changes of scale and aspect (potentially reliant on availability of future image datasets).

5.3.3 Informed action labelling

The use of verbs and verbal phrases in labelling actions in moving images remains a challenge for computer vision models. The nature of crowdsourced training data, especially the paucity of lexicon, has resulted in the generic application of basic action verbs (walking, running, sitting, dancing). Furthermore, some interesting anomalies occur in verb attribution with, for example, a close-up of dancing legs being identified as cutting scissors. Again, context will resolve some of these issues. However, more training data and a greater number of examples of each action (e.g. where a character is not only *walking* but rather *staggering*, and *skipping* or *hopping* rather than merely *running*) are needed to make significant improvements. Human annotators intuitively realise that a person walking erratically after leaving a nightclub is likely to be *staggering* rather than *limping*, and that a child with a sports injury is probably *limping* and not *skipping*. We gather these notions from context and life experience, and at present the machine models are lacking in both.

Action Labelling	Advance action identification to improve specificity, using topic detection techniques to draw on a richer and more contextually relevant lexicon.

5.3.4 Temporal sequencing

In the process of creating human-derived annotations we exercise our innate ability to recognise lexical, textual and visual cues that suggest the passing of time across a narrative. This may occur intra-diegetically as day turns to night, or in an extra-diegetic narration where we are informed that the next scene takes place “10 years later”. Even in instances where narrative temporality is non-linear (e.g. *500 Days of Summer*), our understanding of time passing contributes to the way we construct a narrative arc. Lexical terms found in film dialogue can act as a shorthand for this purpose (*later, next, tomorrow, tonight, yesterday*) and, matched with visual indicators such as changes in location or lighting, are cues the human is programmed to notice. These must be trained into computer vision models to aid temporal sequencing, and form cohesive ties between shots and scenes. Cohesive ties gleaned from the re-occurrence of certain characters or objects (e.g. the same house being filmed repeatedly as shorthand to introduce a particular family) are dependent on accurate computer vision results, but once improved would assist greatly with narrative sequencing and the interconnectedness of plotlines.

Sequencing	Collate cues drawn from dialogue and visual sources to establish basic temporal shifts in narrative (e.g. next day, later, that night).

5.3.5 Establishing narrative saliency

Machines have not achieved a level of sophistication where they can think and feel like human beings. In our previous deliverables we have discussed the importance of filtering the vast amount of information relayed through moving images to extract that which is narratively salient (D5.1, D5.2). Humans are 'programmed' to seek relevance for the purposes of meaning-making, a fact that has received considerable attention in relation to audiovisual data (with studies on aspects such as eye-tracking and focalisation). One step towards training computer models to this end would be to incorporate topic modelling and scene detection into the machine model, as this will determine the main thrust of the narrative and can be used to improve caption generation by limiting the lexical choices to words ('synsets' or 'bag of words') drawn from conceptually and semantically related resources (WordNet, 2020; https://wordnet.princeton.edu/).

Saliency	Combine topic detection techniques with Wordnet libraries to build saliency checking mechanisms into machine description creation.

5.3.6 Sensitivity to the narrative paradigms of storytelling

One aspect of storytelling that presents a tougher challenge for computer modelling workflows is human sensibility to the conventions of narrative exposition. From an early age, human beings learn that certain schemata and paradigms are used for recounting stories of particular genres. It starts with nursery rhymes and fairytales (good versus evil, moral lesson, good wins out over evil), continues with the classic novel (the opening gambit or 'hook', characterisation of protagonists, linear plotting, central conflict, progression to resolution of conflict); and extends to the Hollywood blockbuster (action-packed, multiple conflicts, highly dramatised, weaker characterisations, spectacle more important than plot). Frameworks like this create expectations in consumers, meaning that when we engage with these forms of storytelling we are conditioned to find ways to unravel the narrative to fit the traditional mould.

Machine models have not yet segmented multimodal material into narrative components, currently only applying segmentation at the shot-change level (see above for developments which are currently changing this state of affairs). However, if it were possible to use SG principles to train the machine to recognise the shift from *setting* to *initiating event*, for instance, this would initiate engagement with narrative sequencing. Where it is possible to recognise that protagonists A and B appear in the opening shots of the scene, and that protagonist A goes on to initiate an event which elicits a response in protagonist B, it becomes

possible to link the two scenes using pronominalisation and temporal references. The result immediately becomes more 'human-like' and synthesizes a kind of understanding that narrative occurs on a continuum from meeting the characters through witnessing their challenges to seeing how matters conclude. Story grammar has enabled us to explore basic concepts of plot exposition in the context of the MeMAD500 film corpus, and to test the assumption that all narrative has an underlying structure that is broadly predictable and might be trained into the models driving high level neural networks.

Storytelling	Develop computer vision and audio processing (ASR-based) models to replicate basic narrative shifts between story milestones (*setting, initiating event, internal response, attempt, consequence, reaction*).

5.3.7 Summary

The story grammar methods we have been testing over recent months have the potential to inform machine segmentation, moving from the current system of shot-change segmentation to something that is more narratively relevant and leads to greater cohesion in automated captions. In many narrative scenarios, including news footage, actualité and magazine programming, documentaries and fictional productions, we would suggest that the narrative shift between *setting* and *initiating event* and between *plan/attempt* and *consequence/reaction* are likely to be the most readily discernible to the human audience and therefore the best candidates for early attempts to train the machine in the same skillset. In particular, the shift from *setting* to *initiating event* is often visually depicted as a shift from generic establishing shots ('in the woods', 'at the mall') to the particular ('a man climbs a tree', 'the woman is buying a dress'). There is likely to be a commencement of dialogue when the initiating event occurs, although this is not always the case, and there is also likely to be action in the sense of moving people or objects. With sufficient training data, and further development of feature extraction methods, it should be possible to train the machine to make a reasonable attempt at detecting this early narrative shift. It is then a short step to producing video descriptions which acknowledge this shift ('In a shopping mall'….. 'the woman enters a shop in the mall and speaks to the assistant').

All of this is, to some extent, conjecture. However, for the remainder of the project we will continue to work with partners on improving the editing potential of film segments created using shot-detection methods on the *Flow* platform. Video descriptions of this type are most appropriate for use in archive retrieval contexts, where there may be commercial gains to be made from either short-circuiting current captioning workflows, or creating brief factual descriptions where none were previously available. These tasks are more closely aligned to our 'content description' style of human annotations, where interpretation and elaboration are sacrificed for a more literal, but certainly still useful, record of events.

6 Conclusion

6.1 General Conclusions

This work package project team (WP5) set out to model human methods for creating video descriptions of moving imagery and to use this information to inform future machine modelling. We explored the complexities of human meaning-making using traditional mental modelling methods (Johnson-Laird, 1983), drawing on principles of relevance (Sperber & Wilson, 1995) and coherence (Braun, 2011, 2016; Vandaele 2012) in the context of sequenced narrative. Our quantitative analysis of the lexicon of human video content describers, and that of machine-derived descriptions, highlighted the level of lexical paucity and lack of syntactic sophistication present in current computer captioning models. Appreciating that an analysis of audio description would not provide us with the answers we sought (see below), we created a corpus of 500 film extracts which were annotated in three work streams (*key elements, content descriptions* and *event narration*) in order to examine the layers of human meaning-making and to compare these with computer description iterations.

In the latter phases of our study we explored a smaller sub-corpus of film extracts (20 clips) in order to understand at a more granular level the way human beings take a series of discrete events displayed in a prescribed sequence, and construct a broader narrative with a purposeful start, middle and end. Furthermore, we have adapted the types of story grammars traditionally applied to investigating memory and recall (Stein & Nezworski 1978; Nezworski, Stein & Trabasso 1982; Mandler & Johnson 1977, 1980; Lehnert 1981) and the early examination of computation methods (Rumelhart, 1975) to establish the way in which viewers discern key milestones in storytelling, and compile them in a manner that gives shape to the narrative arc. From this, we found synergies with the development of computer model-building and suggested first steps to move the computer model into a more 'human-like' phase of narrative analysis, starting with the small shifts between narrative milestones such as *setting* and *initiating event*, or *plan/attempt* and *consequence.* Our analysis of visual and audio cues leading to these shifts suggested that computer vision methods should be combined with topic detection and associated lexical *synsets* to optimise outcomes, although there may also be value in weighting towards visual cueing (section 4.4).

One result of our study that was realised early in the project workflow is that human-generated audio description is not particularly helpful as a means of training the machine to detect cues for storytelling in a human-like manner. As a method of intermodal transfer, audio description scripting is far more complex a process than simply stating what is visible on screen. It demands skills of observation, audio extraction, prioritisation, simplification, condensation, omission and deep narrative immersion. These processes require high-level cognitive skills that call upon interpretation, executive function and life experience which are only available to computer models in a very limited way at present.

Therefore, for now, we have pitched our findings at the heart of content description for video cataloguing and retrieval. Our concept of narration and sequencing applies equally to material such as documentaries and news bulletins as it does to film and television drama. Where there is a story to be told, SG can be used to mark the milestones that move the story towards its denouement. At the lowest level, Rumelhart's (1975) notion of stories comprising *event*, *goal* and *outcome* helps us to pinpoint the crux of the action (*goal*) and in doing so, compartmentalise everything that comes before this moment as *event* and what follows as *outcome.* For video retrieval, this would drive the archivist to the most salient moments more rapidly than at present, where a search through the whole edit to find the main event would be necessary.

We would therefore recommend that story grammars are used to inform the next steps in computer modelling for archive retrieval. Automation of audio description is still some way off, and would require significantly more advanced neural networks than are presently available.

6.2 Looking to the Future

Starting with the narrative shift from *setting* to *initiating event*, computer vision models would need to combine feature extraction and audio-assisted methods to identify the transition. Films and documentaries frequently use 'establishing shots' as a way to present the preamble to a scene. The shift from establishing shot to the first narrative activity – whether an action or a line of dialogue – is most frequently accompanied by a change of shot. This may be the move from exterior to interior, or a zoom feature moving from wide-angle to close-up. As humans, we see these as cues for the salient narrative to begin. Combining visual feature extraction with ASR and through this, topic detection, the associated data should in most cases serve to identify this first narrative shift.

Similarly, computer models built to detect the final scene in a narrative sequence, perhaps based on the continuity of either characters present or scenic characteristics/location, could be retro-fitted to establish the beginning of this same narrative sequence, i.e. when the characters or location first appeared. This segment would then be labelled as *outcome* (Rumelhart, 1975), or in a more complex approach, *consequence* and *reaction* (Stein & Nezworski, 1978). In this way, the computer model would be trained to compile narratively sequenced segments as a first step towards meaningful storytelling.

Within WP5, our work with story grammars will continue for the remaining months of the project. We propose to extend our narrative segmentation experiment to a wider audience and test the findings reported here which were limited to just two annotators.

As a final word, it should be noted that the role of training data in the future development of computer vision models cannot be overstated. The quantity and, particularly, the quality of data currently available is not sufficient to make the seismic shift from basic and unreliable

discrete caption generation to something more dependable and narratively meaningful. Developing image data capture from reliable sources needs to be a top priority going forward.

In the meantime platforms such as *Flow*, developed to allow human and machine interaction and a 'sharing of the load' in creating and editing meaningful video descriptions, afford the opportunity to learn more about human meaning-making *in situ* while also offering a way to collect significant amounts of quality data which can be used for future machine training.

7 References

500 Days of Summer (2009) Directed by M. Webb [Feature Film]. USA: Dune Entertainment.

An Education (2009) Directed by L. Scherfig [Feature Film]. UK: BBC Films.

Appose, A. and Karuppali, S. (2018) 'Decoding the macrostructural form of oral narratives in typically developing children between 6–11 years of age: using story grammar analysis'. *Online Journal of Health Allied Sciences*, 17(1), p.12. Available at: https://www.ojhas.org/issue65/2018-1-12.html (Accessed: 07 October 2020).

Banney, R. M., Harper-Hill, K., & Arnott, W. L. (2015), 'The Autism Diagnostic Observation Schedule and narrative assessment: Evidence for specific narrative impairments in autism spectrum disorders.' *International Journal of Speech-Language Pathology,* 17(2), pp. 159–171. https://doi.org/10.3109/17549507.2014.977348 (Accessed: 07 October 2020).

Being Julia (2004) Directed by I. Szabó [Feature Film]. UK: Sony Pictures Classics.

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

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

Brown, D., Brown, E., Lewis, C. and Lamb, M. (2018) 'Narrative skill and testimonial accuracy in typically developing children and those with intellectual disabilities'. *Applied Cognitive Psychology*, 32(5), pp. 550-560.

Bruce Almighty (2003) Directed by T. Shadyac [Feature Film]. USA: Spyglass Entertainment.

Casino Royale (2006) Directed by M. Campbell [Feature Film]. UK: Eon Productions.

Cozendey, S. and da Piedade Costa, M. (2016) 'The Audio Description as a Physics Teaching Tool', *Journal of Research in Special Educational Needs,* 16(1), pp. 1031-1034.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A. and Meignier, S. (2018) 'An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality'. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 15th- 20th April, Alberta, Canada.

Extremely Loud and Incredibly Close (2011) Directed by S. Daldry [Feature Film]. US: Scott Rudin Productions.

Goal (2005) Directed by D. Cannon [Feature Film]. UK: Touchstone Pictures.

Hayes, D. and Birnbaum, D. (1980) 'Preschoolers' retention of televised events: Is a picture worth a thousand words?'. *Developmental Psychology*, 16(5), pp. 410-416.

Hayes, D.S. and Kelly, S.B. (1985) 'Sticking to syntax: The reflection of story grammar in children's and adults' recall of radio and television shows'. *Merrill-Palmer Quarterl,* pp. 345-360.

He, K., Zhang, X., Ren, S., & Sun, J. (2016) 'Deep Residual Learning for Image Recognition'. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778. Available at: https://arxiv.org/abs/1512.03385 (Accessed: 07 October 2020).

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

Johnny English Reborn (2011) Directed by O. Parker [Feature Film]. UK: Studio Canal.

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

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

Lehnert, W.G. (1981), 'Plot units and narrative summarization'. *Cognitive science*, 5(4), pp. 293-331.

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

Mandler, J.M. (1982) 'Some uses and abuses of a story grammar'. *Discourse Processes*, 5(3-4), pp. 305-318.

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

Mandler, J.M., & Johnson, N.S. (1980) 'On throwing out the baby with the bathwater: A reply to Black and Wilensky's evaluation of story grammars'. *Cognitive Science*, 4(3), pp. 305-312.

Memoirs of a Geisha (2005) Directed by R. Marshall [Feature Film]. USA: Columbia Pictures.

Nezworski, T., Stein, N.L. and Trabasso, T. (1982) 'Story structure versus content in children's recall'. *Journal of Verbal Learning and Verbal Behavior*, 21(2), pp. 196-206.

Park, J.S., Rohrbach, M., Darrell, T. and Rohrbach, A. (2019) 'Adversarial inference for multi-sentence video description'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6598-6608.

Pretty Woman (1990) Directed by G. Marshall [Feature Film]. USA: Touchstone Pictures.

Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L-J. (2017) 'Deep Reinforcement Learning-based Image Captioning with Embedding Reward'. Available at:https://arxiv.org/abs/1704.03899 (Accessed: 07 October 2020).

Rumelhart, D.E. (1975) 'Notes on a schema for stories'. In: *Representation and understanding*, pp. 211-236. Morgan Kaufmann.

Rumelhart, D.E. (1980) 'On evaluating story grammars'. *Cognitive Science*, 4(3), pp. 313-316.

Rumelhart, D., and Ortony, A. (1977) 'The representation of knowledge in memory', in R. Anderson, R. Spiro, and W. Montague (Eds.), *Schooling and the acquisition of knowledge*. Hillsdale, N.J.: Erlbaum.

Sex and the City (2009) Directed by M. Patrick King [Feature Film]. USA: New Line Cinema.

Singer, H. and Donlan, D. (1982) 'Active comprehension: Problem-solving schema with question generation for comprehension of complex short stories'. *Reading Research Quarterly*, pp. 166-186.

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

Stein, N.L., Glenn, C.G. and Freedle, R. (1979) 'New directions in discourse processing'. *Norwood, NJ: Ablex*.

Stein, N.L. and Nezworski, T. (1978) 'The effects of organization and instructional set on story memory'. *Discourse processes*, 1(2), pp. 177-193.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D. Vanhoucke, V. and Rabinovich, A. (2015) 'Going Deeper with Convolutions', *Proceedings of the IEEE conference 2015 on Computer Vision and Pattern Recognition.* Available at: https://arxiv.org/abs/1409.4842 (Accessed: 07 October 2020).

The Aviator (2004) Directed by M. Scorsese [Feature Film]. US: Forward Pass.

The Devil Wears Prada (2006) Directed by D. Frankel [Feature Film]. US: Fox 2000 Pictures.

The Forgotten (2004) Directed by J. Ruben [Feature Film]. US: Revolution Studios.

The Guardian (2006) Directed by A. Davis [Feature Film]. US: Touchstone Pictures.

The Help (2011) Directed by T. Taylor [Feature Film]. US: DreamWorks Pictures.

The King's Speech (2010) Directed by T. Hooper [Feature Film]. UK: UK Film Council.

The Matador (2011) Directed by R. Shepard [Feature Film]. US: Miramax.

The Social Network (2010) Directed by D. Fincher [Feature Film]. US: Columbia Pictures.

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

Whalon, K., Henning, B., Jackson, E. and Intepe-Tingir, S. (2019) 'Effects of an adapted story grammar intervention on the listening comprehension of children with autism'. *Research in developmental disabilities*, 95.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. (2006) 'ELAN: a professional framework for multimodality research'. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 1556-1559.

Wordnet.princeton.edu. 2020. *Wordnet | A Lexical Database For English*. [online] Available at: https://wordnet.princeton.edu/ (Accessed: 7 October 2020).