Braun, Sabine; Starr, Kim; Delfani, Jaleh; Tiittula, Liisa; Laaksonen, Jorma; Braeckman, Karel; Van Rijsselbergen, Dieter; Lagrillière, Sasha; Saarikoski, Lauri

When Worlds Collide: AI-Created, Human-Mediated Video Description Services and the User Experience

Published in:
HCI International 2021 - Late Breaking Papers: Cognition, Inclusion, Learning, and Culture

DOI:
10.1007/978-3-030-90328-2_10

Published: 01/07/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Braun, S., Starr, K., Delfani, J., Tiittula, L., Laaksonen, J., Braeckman, K., Van Rijsselbergen, D., Lagrillière, S., & Saarikoski, L. (2021). When Worlds Collide: AI-Created, Human-Mediated Video Description Services and the User Experience. In C. Stephanidis, D. Harris, W.-C. Li, D. D. Schmorrow, C. M. Fidopiastis, M. Antona, Q. Gao, J. Zhou, P. Zaphiris, A. Ioannou, A. Ioannou, R. A. Sottilare, J. Schwarz, & M. Rauterberg (Eds.), HCI International 2021 - Late Breaking Papers: Cognition, Inclusion, Learning, and Culture: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings (pp. 147-167). (Lecture Notes in Computer Science; Vol. 13096). Springer. https://doi.org/10.1007/978-3-030-90328-2_10

This is a copy of the pre-final version of:

Braun, S. *et al.* (2021). When Worlds Collide: AI-Created, Human-Mediated Video Description Services and the User Experience. HCI International 2021: Cognition, Inclusion, Learning, and Culture. HCII 2021. Lecture Notes in Computer Science, vol 13096. Springer, Cham. https://doi.org/10.1007/978-3-030-90328-2_10

When Worlds Collide: AI-Created, Human-Mediated Video Description Services and the User Experience

Sabine Braun[1], Kim Starr[1], Jaleh Delfani[1], Liisa Tiittula[2], Jorma Laaksonen[3], Karel Braeckman[4], Dieter Van Rijsselbergen[4], Sasha Lagrillière[5], and Lauri Saarikoski[5]

1 University of Surrey, Guildford GU2 7XH, UK

2 University of Helsinki, Yliopistonkatu 4, 00100 Helsinki, Finland

3 Aalto University, 02150 Espoo, Finland

4 Limecraft, Sint-Salvatorstraat 18b/301, 9000 Gent, Belgium

5 YLE, Media House Uutiskatu 5, 00240 Helsinki, Finland

Abstract. This paper reports on a user-experience study undertaken as part of the H2020 project MeMAD ('Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy'), in which multimedia content describers from the television and archive industries tested *Flow*, an online platform, designed to assist the post-editing of automatically generated data, in order to enhance the production of archival descriptions of film content. Our study captured the participant experience using screen recordings, the User Experience Questionnaire (UEQ), a benchmarked interactive media questionnaire and focus group discussions, reporting a broadly positive post-editing environment. Users designated the platform's role in the collation of machine-generated content descriptions, transcripts, named-entities (location, persons, organisations) and translated text as helpful and likely to enhance creative outputs in the longer term. Suggestions for improving the platform included the addition of specialist vocabulary functionality, shot-type detection, film-topic labelling, and automatic music recognition. The limitations of the study are, most notably, the current level of accuracy achieved in computer vision outputs (i.e. automated video descriptions of film material) which has been hindered by the lack of reliable and accurate training data, and the need for a more narratively oriented interface which allows describers to develop their storytelling techniques and build descriptions which fit within a platform-hosted storyboarding functionality. While this work has value in its own right, it can also be regarded as paving the way for the future (semi)automation of audio descriptions to assist audiences experiencing sight impairment, cognitive accessibility difficulties or for whom 'visionless' multimedia consumption is their preferred option.

Keywords: Computer vision, Video description, Content description, Audiovisual translation, Archive retrieval, Audio description, Media accessibility

1 Introduction

1.1 Background

In the ongoing debate about the value of AI in human-dominated workstreams there are clearly tasks the human completes with greater compassion, empathy, subtlety and contextualisation than a machine, and these are skills which remain difficult to train into

computer models (e.g. one which automates audio description or similar video description services). Yet the AI machine, programmed to operate dispassionately and with algorithmic efficiency, is capable of producing large volumes of data in a fraction of the time it would take human operatives (e.g. processing computer vision training data). While each of these methods offers benefits, they also present different challenges. Can automated video description ever match the expectations of audiovisual content creators and editors? Can human endeavour alone keep pace with the proliferation of new media resources requiring description? To what extent is quality negotiable in return for increased volumes and speed of output? And, most importantly, where do media access and the media consumer fit into the picture?

1.2 Study Aims and Structure

In the European MeMAD project (grant no. 780069), our primary focus has been on **developing semi-automated video description models** which replicate, as far as possible, the work of human describers of audiovisual content [1, 2]. This has been achieved using computer vision modelling, theories of human engagement with multimodal narrative, and the integration of machine-generated data within an editing platform, *Flow*, which draws together machine descriptions, named-entity recognition, metadata, transcriptions and translation services (Fig. 1). While the eventual aim is to **create a methodology** for achieving automated (or semi-automated) audio descriptions of high volume, low value media artefacts such as social media streams, it became clear early on in the MeMAD project that this goal is currently unattainable. Given the present level of sophistication achieved with machine-generated video captions [3–5], a more pragmatic approach was taken to produce baseline automated video descriptions and other metadata, which could be made available through an online editing platform. Human operatives would then use the machine-generated data as a starting point to create descriptions of archive film resources for future access and retrieval. As the accuracy of machine descriptions improves in the future, we anticipate the post-editing process will become less onerous for human content describers, freeing them to concentrate their efforts on the highest value audiovisual artefacts in their collections.

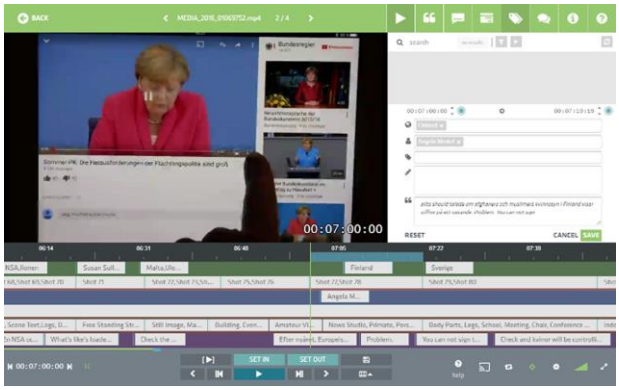

Fig. 1. Flow platform editing mode

This process has the potential to lead to the development of a system for the **semiautomation for audio description**, although suffice it to say this would require a seismic jump in functionality from the current state of affairs. Firstly, it would be necessary to integrate audio cue processing (using automatic speech recognition and topic detection) into the automated caption generation process, followed by identification of the available hiatuses in the original audio track, and subsequently, the application of text condensation techniques to match

automated audio description scripts with gaps in the original audio track. Finally, a data prioritisation system or metric would have to be applied to order the narrative saliency of visual information. All of these complex audiovisual tasks are integral to the successful delivery of human-based audio description.

Within the above context, this paper reports the methodology and results of a recent study undertaken to test user experience on the *Flow* video description editing platform [6] with participants from the television and film archive industries. In particular, consideration was given to the commercial value of using an interface which offers a range of data and metadata from which the user may select and discard options in line with any given editing brief. In doing so, questions were raised about the **impact of semiautomation on workflows**, **data reliability**, and most importantly from the perspective of the end-consumer, **quality of outputs.**

Since our aim was to **assess the workflows** involved in editing automated and prepackaged data within a platform environment, the quality of the machine descriptions and the users' output were not evaluated, nor did we measure or evaluate the speed of production. This was for several reasons: the novelty of the tool for users (lack of familiarity given that the users were new to this particular editing environment); our aim to include participants from different countries, which created some language barriers (the machine video captions were available only in English, whilst participants described the content in different languages); the quality of the sample content (i.e. our assumption, based on an earlier analysis of machine-derived video captions [1, 2], was that they are currently not sufficiently accurate or reliable to provide a realistic starting point for a simulation of true-to-life editing).

Having first outlined our research framework and methods, we will then present and extrapolate our results to consider the implications of harnessing human-machine interaction on editing platforms such as *Flow* in the context of future enhancements, as well as the possibility of extending accessibility beyond archive retrieval.

2 Methodology

2.1 Research Design

Due to the novelty of the workflows under evaluation, a mixed-methods approach to data collection was adopted. This comprised three phases: (i) **observation**, via screen recording software, of study participants undertaking hands-on work editing data on the platform; (ii) a validated **questionnaire**, the UEQ [7], used to evaluate user experience, complemented by specific questions about participants' experience of the *Flow* work environment; and (iii) **focus groups** to elicit participants' suggestions for further development of the platform and refinement of the features and functions of the prototype. In light of the Covid-19 pandemic, data collection was conducted through a series of online workshops with participants hosted via Zoom and the *Flow* platform, and chatroom technologies used for facilitating follow-up discussions.

2.2 Participants

Participants formed a convenience sample, with recruitment being conducted through organisations expressing an interest in the work of the MeMAD consortium. They were drawn from the broadcast and media archive industries including television production staff from a number of European broadcasters, and archivists from the Finnish national archive institution. Twenty-three participants were initially recruited to the study, with eighteen completing all three phases (four participants withdrew due to either time constraints or technical issues). Job

descriptions of those whose finished all three phases (n=18) included roles in television production, production coordinators, assistant producers, archive journalists and cataloguers. All participants had some previous experience describing filmed content. Recruitment took place across four countries (Finland, Sweden, Switzerland, Germany) with participant ages ranging from 30 to 69 years; 67% identified as female and 33% as male. The highest educational qualification participants held was a Master's degree (72%) with the remainder holding either first degrees or school leavers' qualifications (28%). They possessed between 1 and 10 + years' experience in the description of AV content in the television or archive industries. In terms of their expertise of working with similar editing platforms, the participants self-identified as expert (N = 2; 11.1%), advanced (N = 2; 11.1%), intermediate (N = 11; 61.1%), novice (N = 2; 11.1%) or inexperienced (N = 1; 5.6%) users of content editing platforms.

2.3 Study Design and Conduct

Participants were invited to an **evaluation workshop** during which they were given a hosted induction to the platform, followed by the opportunity to engage with the prototype in an individual, expert-led, hands-on session. Embedding the evaluation in a training workshop was regarded as the most effective way to introduce participants to the study, providing them a basic familiarity with the tool, while enabling the research team to observe the process and elicit participants' initial views of the platform. As a consequence of the Covid-19 situation, several participants were working from home, using laptops. They were given access to the *Flow* platform through a private broadband connection; Zoom video conferencing software was used for the training workshop. Six workshops were held, each with up to five participants, all following the same pattern. The first segment comprised a 45-min introduction to the project and platform **induction** by the MeMAD project team including the software developers. Time was given for a brief Q&A. This was followed by a 45-min to one-hour **phase of individual work**, using the Zoom breakout rooms. Participants were given access to five short video clips, a mixture of both contemporary and heritage material (one clip was used as a "warm-up" exercise) and briefed on the content description task(s). To observe the video describers' approach to the content description task and their interaction with the prototype platform, participants were asked to share their screen in Zoom during the hands-on session. These sessions were video-recorded. The focus groups were also video-recorded using the Zoom platform. Participants worked at their own pace, according to experience and technical competence, with the result that the number of annotation tasks completed ranged from between one and five video clips per person. Three participants completed four clips, six participants completed three clips, a further six participants completed two clips; three participants worked on just one clip. Technical support was available to participants in Zoom breakout rooms throughout the hands-on session. After completion of the set tasks, a **questionnaire** was administered to elicit the participants' views on working with the platform. This was followed by a **focus-group** discussion, lasting between 33 and 71 min across the six workshops.

2.4 Data Collection

The questionnaire consisted of four sections: basic demographics; the standard User Experience Questionnaire (UEQ); and two additional sections comprising questions specific to the current study. The UEQ [7] is a widely adopted data collection instrument, used to elicit users' impressions, feelings and attitudes towards a range of interactive products similar to the

Flow platform. It comprises twenty-six 7-point Likert-type questions, intended to measure usability and user experience across six dimensions. (Table 1).

Although the *Flow* platform is currently at an early stage of development, all components of the UEQ were used in this evaluation. The main focus was on **usability**, however, we also sought to elicit participants' views on **attractiveness** and **user experience** to inform future development. Interpretation of the outcomes for these components was conducted in a way that was mindful of the prototype nature of the platform.

The second part of the questionnaire contained two further sets of 7-point Likert-type questions, one relating to the **work environment** and the other eliciting users' preferences in relation to the particular **features and functions** of the *Flow* prototype. The work environment section interrogated participants' impressions of **process and workflows** (e.g. *"I felt comfortable working in this environment"*) and their perception of the opportunity to create more efficient or effective descriptions using the platform (e.g. *"I feel that the environment has helped me to produce good descriptions"*). The **preferences** section investigated the participants' attitudes towards specific characteristics of the prototype (e.g. *"The 'adding a content description' feature was efficient/functional", "The timeline lane showing places, persons, tags was useful"*).

Table 1. Dimensions of the UEQ (http://www.ueq-online.org)

Grouping	Dimension	Explanation
Overall	Attractiveness	Overall impression of the product. Do users like or dislike it?
Usability	Perspicuity	Is it easy to get familiar with the product and to learn how to use it?
	Efficiency	Can users solve their tasks without unnecessary effort? Does it react fast?
	Dependability	Does the user feel in control of the interaction? Is it secure and predictable?
User experience	Stimulation	Is it exciting and motivating to use the product? Is it fun to use?
	Novelty	Is the design of the product creative? Does it catch the interest of users?

2.5 Data Analysis

The standard sections of the UEQ **questionnaire** were analysed using an instrument integral to the package. This quantifies basic data on user experience by comparing participants' responses against a benchmark dataset, consisting of metrics from over 14000 participants evaluating more than 280 products (e.g. business software, web pages, online shops, social networks). From these studies, benchmarked mean scores for each of the six dimensions of the UEQ are supplied. The UEQ analysis was complemented by a statistical analysis of the overall experience questions, and the questions relating to specific features and functions, all of which were particular to this study. Answers to open-ended questions were analysed qualitatively with a focus on user preferences. To complement this data, **focus group discussions** were thematically analysed to compare and contrast participants' perceptions of the prototype. Wherever possible, references in the discussion were related to specific instances in the observed **hands-on sessions** and to the questionnaire responses of the respective participants. Our purpose was to use the focus groups to make a more granular study of the observed actions and questionnaire responses obtained in the earlier phases of the study.

2.6 Ethical Considerations

The study was approved by the University of Surrey Ethics Committee (Reference number: FASS 20–21 014 EGA). Participants were recruited from four nations, with most participants indicating that they were comfortable participating in English. The questionnaire was made

available in English and Finnish to accommodate different language backgrounds. Focus groups were held in both Finnish and English.

3 Results

3.1 User Experience Questionnaire (UEQ)

Generic UEQ Questions. As illustrated in Table 1, the 26 items in the UEQ are grouped into six 'dimensions' representing usability and user experience. The results for each individual metric are presented as mean scores in Table 2; the scores for each of the six dimensions are shown in Table 3. UEQ scores range between −3 (extremely bad) and +3 (extremely good).

Table 2. UEQ mean scores for individual questions

Item	Mean	Variance	Std. Dev.	No.	Left	Right	Dimension
1	1.00	1.18	1.08	18	Annoying	Enjoyable	Attractiveness
2	1.06	1.23	1.11	18	Not understandable	Understandable	Perspicuity
3	1.17	0.85	0.92	18	Creative	Dull	Novelty
4	1.06	1.47	1.21	18	Easy to learn	Difficult to learn	Perspicuity
5	1.39	0.72	0.85	18	Valuable	Inferior	Stimulation
6	1.17	0.62	0.79	18	Boring	Exciting	Stimulation
7	1.94	0.41	0.64	18	Not interesting	Interesting	Stimulation
8	0.72	1.39	1.18	18	Unpredictable	Predictable	Dependability
9	1.28	1.74	1.32	18	Fast	Slow	Efficiency
10	1.17	2.03	1.42	18	Inventive	Conventional	Novelty
11	1.28	0.45	0.67	18	Obstructive	Supportive	Dependability
12	1.00	2.12	1.46	18	Good	Bad	Attractiveness
13	0.83	1.56	1.25	18	Complicated	Easy	Perspicuity
14	1.44	0.73	0.86	18	Unlikable	Pleasing	Attractiveness
15	0.94	0.41	0.64	18	Usual	Leading edge	Novelty
16	1.17	0.62	0.79	18	Unpleasant	Pleasant	Attractiveness
17	0.89	1.16	1.08	18	Secure	Not secure	Dependability
18	1.39	1.31	1.14	18	Motivating	Demotivating	Stimulation
19	0.94	1.35	1.16	18	Meets expectations	Does not meet expectations	Dependability
20	1.06	1.23	1.11	18	Inefficient	Efficient	Efficiency
21	1.00	1.53	1.24	18	Clear	Confusing	Perspicuity
22	1.33	0.94	0.97	18	Impractical	Practical	Efficiency
23	1.11	1.16	1.08	18	Organized	Cluttered	Efficiency
24	1.11	0.81	0.90	18	Attractive	Unattractive	Attractiveness
25	1.33	1.41	1.19	18	Friendly	Unfriendly	Attractiveness
26	1.44	0.85	0.92	18	Conservative	Innovative	Novelty

Table 3. UEQ 'Dimensions'

Dimension	Mean	Variance	Std. Dev.
Attractiveness	1.18	0.56	0.75
Perspicuity	0.99	1.09	1.04
Efficiency	1.19	0.86	0.93
Dependability	0.96	0.65	0.81
Stimulation	1.47	0.37	0.61
Novelty	1.18	0.57	0.76

However, the UEQ developers note that mean scores of above +2 or below −2 are unlikely to be observed due to a tendency for respondents to avoid both extremes when presented with a Likert scale survey. According to the UEQ development team, values between −0.8 and +0.8 represent a neutral evaluation of the corresponding item or dimension, values >0.8 represent a positive evaluation and values <0.8 indicate a negative evaluation. In line with this assessment, the mean scores recorded for both individual items and the six dimensions suggest that participants were rating the platform positively. The UEQ's rubric for estimating required sample size for generalisability, based on the level of precision E (i.e. difference between true scale mean in the population and estimated scale mean from the sample) and the standard deviation (in the sample), suggests that our sample size was large enough for E = 0.5 and for an error probability P = 0.05 for perspicuity and efficiency and P = 0.01 for attractiveness, dependability, stimulation, novelty.

In general, it can be observed that the mean scores for all six 'dimensions' fall above the >0.8 threshold identified by the UEQ development team as indicating positive feedback. Despite *Flow* still being in the prototype phase, *attractiveness* and the two *user experience* dimensions (*stimulation, novelty*) were evaluated very positively. Interestingly, the highest scores *(mean = 1.47)* relate to *stimulation* which, in the context of an industry where automation is often viewed with suspicion and perceived as a potential threat to job satisfaction, is highly encouraging. *Attractiveness (mean = 1.18), efficiency (mean = 1.19)* and *novelty (mean = 1.18)* also score strongly. Unsurprisingly, for a new platform with a sharp learning curve, *perspicuity (mean = 0.99)* registered more modest (though still positive) scores, and dependability *(mean = 0.96)* while also an encouraging score, suffered from the vagaries of remote connectivity and, to some extent, the lack of reliability still evident in machine-generated descriptions.

In order to contextualise these scores, the study's UEQ results were benchmarked against an industry reference dataset made available by the UEQ developers. As noted above, this dataset is active and growing, but currently includes over 14000 questionnaire responses from 280 studies derived from a broad selection of interactive and digital product research studies. The benchmarked results for *Flow* are shown in Fig. 2. The classifications used in benchmarking are 'excellent' (meaning that the results are in the range of the 10% best results in the benchmark dataset), 'good' (10% of the results in the benchmark dataset are better and 75% are worse), 'above average' (25% of the results in the benchmark dataset are better and 50% are worse), 'below average' (50% of the results in the benchmark dataset are better and 25% are worse) and 'bad' (in the range of the 25% worst results).

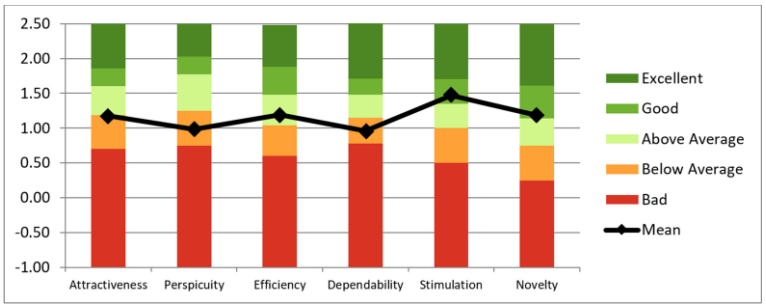

Fig. 2. Benchmarked UEQ scores compared with *Flow* mean scores

Figure 2 shows mean scores for *Flow* (marked in black) against benchmarked categories (excellent, good, above average etc.). Whilst two dimensions were slightly below the benchmarked average (*perspicuity* and *dependability*), one scored above average (*efficiency*) and two were rated as good (*stimulation, novelty*). The score of *attractiveness (mean = 1.18)*

was good in absolute terms (as good as the score for *efficiency*), but it was marginally lower (0.02) than the benchmarked average for this category, indicating that our evaluators found the platform attractive, but very marginally less attractive than the average product in the UEQ benchmark dataset. The two user experience dimensions (*stimulation*, *novelty*) benchmarked well against the reference dataset, which might be expected given that *Flow* offers a unique approach to archive development. The possible reasons for the ratings for *perspicuity* and *dependability* were outlined above. Further insights into the participants' perceptions can be derived from the 'Working environment' section of the questionnaire (below).

Working Environment. Participants' Responses regarding their general experience of the working environment – i.e. their interaction with the platform in their set-up, including their computer, workstation and internet connection – are presented as summative scores below (Fig. 3), based on the participants' perceptions of the naturalness in the use of the platform within their work environment, how comfortable they felt working in their environment, the impact that the work environment had on their performance, and whether the environment helped them to produce viable descriptions.

Based on these four questions, and using a 7-point Likert scale, the minimum and maximum scores were 1 and 28 respectively. The overall experience associated with the prototype was scored at M = 12.22 (SD = 3.54).

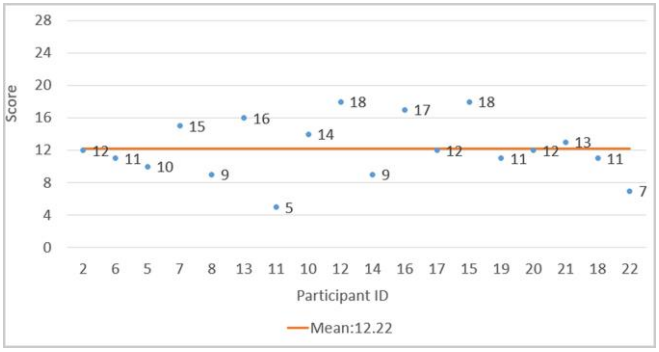

Fig. 3. Experience of working environment

The most positive perceptions were observed in the group registering intermediate experience with similar software platforms (N = 11; M = 12.36, SD = 3.64), the novice group (N = 2; M = 12.50, SD = 4.95) and the expert group (N = 2; M = 12.00, SD = 1.41), while the advanced group's score was lower (N = 2; M = 8.50, SD = 2.12). The participant identifying as inexperienced gave a score of 18, which was the highest score awarded. However, given the small participant numbers per group, these results need to be treated with caution.

The breakdown by years of experience creating content descriptions did not reveal large differences (1–5 years: N = 5, M = 11.2, SD = 2.86; 6–10 years: N = 3, M = 13.33, SD = 7.23; > 10 years: N = 10, M = 12.4, SD = 2.95). Stronger differences emerged in relation to the participants' professional affiliation. Participants from company A (N = 3) and company C (N = 4) scored their overall experience of working with the prototype at M = 15.33 (SD = 3.06) and M = 14.00 (SD = 4.24) respectively, whilst the scores given by participants from company B (N = 4) and D (N = 6) were M = 10.75 (SD = 2.63) and M = 10.67 (SD = 3.61). This is most likely linked to the companies' current workflows but it may in part be explainable by the fact that the organisations present different work environments, i.e. broadcaster vs. national archive.

For example, of the participants working in a **broadcasting environment**, those with intermediate expertise of using similar platforms commented that the tool was quite easy to

handle and enjoyable, but that the combination of not knowing the context of the clips and not being familiar with the prototype interface made the specific task difficult, and that more time than available during the workshop would be needed to become familiar with the platform. Broadcaster-based participants with different levels of expertise (intermediate, expert) reported difficulties adjusting the time code of automatically presegmented segments, where they felt such adjustments were necessary. Furthermore, some individuals reported that text they had entered in the description fields seemed to disappear and had to be re-entered (intermediate, advanced). Other comments revealed problems with using the video player during the description, and problems with clearing data from some fields. Two participants had difficulty with playing the video clips on a Mac computer (expert) and processing the recorded video files from Zoom (intermediate). One participant (expert) thought that there was not enough automatic extraction of metadata to help description, especially with regard to face recognition [8]. Finally, the comments from the broadcaster-based participants also point to another source of difficulty: the participants' working environment, which was at home due to the pandemic. Whilst three of these participants did not have any technical problems, others felt that their (laptop) screen was too small, leading to them not being able to see all features of the interface at the same time.

Participants working in an **archival environment** reported relatively few technical problems. One participant from this group (novice) thought that moving the timeline was difficult, as the content moves while the track head stays in place, which was different from this participant's (limited) own practical experience. Another participant in this group (intermediate) felt that the user interface should communicate a little better, for example, warning the user when a description they produced would not be saved once they move to the timeline. One participant (intermediate) suggested that a list of terms to use could be helpful, depending on the purpose of the description.

Specific Features and Functions of the Prototype. The results of the specific features and functions of the prototype section are presented as measures of central tendency. Based on a 7-point Likert scale, eight of the sixteen items that participants were asked to score (Table 4) were rated average or above.

Table 4. Views about specific features of the *Flow* platform

Q#	Question	Mean	SD	Median	Mode
39	It was easy to access the application	2.67	1.41	2	2
40	The information provided in the application was not too technical	5.50	1.07	5.5	5
41	The 'adding a content description' feature was efficient/functional	3.22	1.31	3	3
42	The timeline was useful for navigating the clip	2.78	1.62	2	2
43	The timeline was useful for creating new content descriptions	3.28	1.97	3	2
44	The timeline zoom and pan was easy to use	3.89	1.59	4	4
45	Selecting a time range using the SET IN / SET OUT buttons was intuitive	3.39	1.60	3	5
46	The places, persons and other suggested tags were useful	3.50	1.21	3	3
47	The suggestion in the 'spoken text' field' was useful	3.44	1.61	3	3
48	The sidebar with existing content descriptions was clear	3.33	0.94	4	4
49	The timeline lane showing places, persons, tags was useful	3.56	1.71	3	3
50	The 'Deep Caption' timeline lane was useful	4.39	1.38	4	4
51	The 'Shots' timeline lane was useful	3.72	1.88	4	4
52	The 'Faces' timeline lane was useful	4.11	1.49	4	4
53	The 'OCR' (text detected in screen) timeline lane was useful	3.72	1.56	4	4
54	The 'Transcript' lane for the language spoken in the clip was useful	3.06	1.68	3	3

While the low scores for *access to the application* (Q39) require further scrutiny, the participants were generally appreciative of the way in which the information was presented on the platform (Q40). The scores for the core function of *adding a content description* (Q41) are average, but interestingly the various *support feeds* offered in the platform to enable the human operator to create (write) the descriptions were all perceived as being useful, especially the various timeline lanes (tiers) showing the *shot segmentation* (Q51), the *automated video captions* (Q50), the results of the *automatic face recognition* (Q52), *text detected in the AV content* (Q53), and the transcript (Q54). The various displays of *tags for persons, places and other features* were also deemed helpful (Q46, Q49) as was the *'spoken text' snapshot*, which highlighted quotes from the *transcript* (Q47). One participant with intermediate experience with editing platforms thought that the tag suggestions and the suggested (automatically generated) video descriptions were by far the most useful material; more useful than the timeline lanes. Another participant with intermediate experience commented that the tool as a whole is useful when there is enough time to learn how to use it and to work at one's own pace, without time pressure.

The perceptions of *working with the timeline* (Q42–45) were more mixed. In the feedback comments, one of the participants (with intermediate experience) noted that s/he did not fully understand how to move on the timeline. An advanced user stated that many points could be made about the timeline, but that its usefulness ultimately depends on the data available to be displayed in the timeline lanes. The feedback garnered in the focus group discussions (see Sect. 3.2) gives more insight into the participants' thoughts about the timeline features.

Further analysis shows that the reactions to the prototype's features and functions varied according to the participants' level of expertise with editing platforms (Fig. 4).

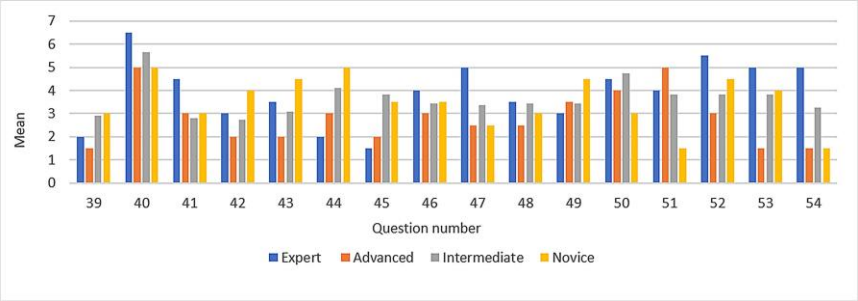

Fig. 4. Features and functions according to level of expertise (1)

Participants identifying as expert users of editing platforms (N = 2) had the most positive views on eight of the sixteen items (Table 5):

Table 5. Participants identifying as experts

Item#	Functionality
40	The information provided in the application was not too technical
41	The 'adding a content description' feature was efficient/functional
46	The places, persons and other suggested tags were useful
47	The suggestion in the 'spoken text' field' was useful
48	The sidebar with existing content descriptions was clear
52	The 'Faces' timeline lane was useful
53	The 'OCR' (text detected in screen) timeline lane was useful
54	The 'Transcript' lane for the language spoken in the clip was useful

Those identifying as beginners (N = 2) had the most positive views on five of the items (Table 6):

Table 6. Participants identifying as beginners

Item#	Functionality
39	It was easy to access the application
42	The timeline was useful for navigating the clip
45	Selecting a time range using the SET IN and SET OUT buttons was intuitive
49	The timeline lane showing places, persons, tags was useful

Furthermore, consistent with the assessment of the working environment, the expert, intermediate and novice groups were more positive in their assessment of the features than the advanced group. However, given the small number of participants in the individual expertise-level groups, the apparent differences need to be treated with caution. A breakdown according to users with higher levels of experience (expert and advanced, N = 4) and lower levels of experience (intermediate, beginner, N = 13) suggests that, with the exception of two technical features of the timeline (44, 45), the participants' perceptions of the prototype's features are relatively consistent (Fig. 5).

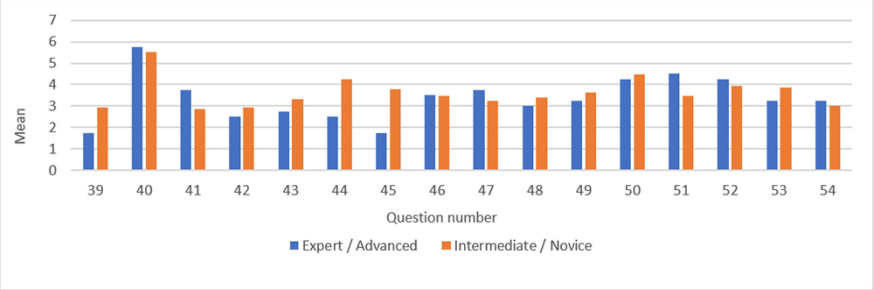

Fig. 5. Features and functions according to levels of expertise (2)

3.2 Focus Groups

Workshop and Prototype Evaluation. The focus groups (FG) largely corroborated the survey findings, revealing participants' **positive overall perceptions** of the workshop and the prototype. Most participants felt that the workshop was a good experience, and they described the prototype as interesting, novel, impressive, handy, intuitive, functional and logical. Consistent with the positive UEQ score for the effort required to learn how to work with the platform (M = 1.1, SD = 1.2), only a small number of participants reported in the FG that it was difficult to learn how to work with it. Some participants reported that they had **technical difficulties**, especially at the beginning of the hands-on session. Although these difficulties were resolved quickly, they may, in part, explain the low score for ease of access to the application in the specific features and functions section of the questionnaire (Q39: M = 2.67, SD = 1.41). In addition, the small screen size of the laptops that some participants used interfered with viewing the whole *Flow* platform at a glance. A recurrent theme across the focus groups was **familiarisation**. Several participants expressed regret at not having been given more time in the workshop to familiarise themselves with the tool. Some participants felt that this had made the evaluation somewhat difficult. Yet, participants who had seen earlier versions of the automatically generated metadata noted a clear progress, for example, in speech recognition.

Creating Content Descriptions from Machine-Generated Data. Participants' comments highlighted that their companies' archive systems are undergoing change (e.g. through the introduction of speech technology), and that they would welcome a tool with the functionalities offered by the *Flow* platform prototype, as it corresponds to **new ways of working** in the broadcast industry. The participants acknowledged the potential of the **machine-enhanced human workflow** supported by the prototype, i.e. the creation of content descriptions based on machine-generated metadata/tags and video captions. Most participants felt that this workflow could, in principle, facilitate the content describers' task, for example by providing a starting point for a description and helping to increase the consistency of the descriptions. Interestingly, some of the highly experienced participants explained that they are so used to looking at the video footage in their normal practice that they initially ignored the automatic video captions. However, on closer inspection of the captions they felt that the captions could be helpful for **understanding unclear or highly unfamiliar content**, and that reading the captions enabled them to **identify information** they had missed in the video footage.

One participant felt that the automatic creation and ingestion of metadata is particularly relevant for legacy content without metadata, whilst for new productions, the production team would normally create basic metadata today (i.e. characters, places, keywords). In the case of new productions, metadata can be directly extracted from a planning/programming system, or this type of metadata can be combined with automatically created information.

As expected, however, the participants were critical of the **quality of the metadata and video captions** used in the evaluation, noting that this data was often flawed and that it is not possible to verify the information. Some participants thought that the video captions were more useful for the samples of contemporary video footage, and less effective for the legacy material. One participant pointed out that erroneous automated captions, which require much amendment, could be more trouble than they are worth. Another highlighted the potentially dire consequences of erroneous descriptions based on erroneous captions in some contexts. A further participant wondered at a machine's capability of identifying **salient or relevant information** in a video scene and contended that this requires human interpretation. Finally, some participants felt that the descriptions were **too detailed or fragmented** and that human describers would normally describe AV content at a higher level of abstraction. However, others pointed to the need for detail, explaining that from a search and retrieval perspective it would be more useful to be able to retrieve instances of "a playing child" rather than "a child". Options of this nature will, inevitably, reflect the video description protocols applied in each individual's place of work.

Overall Positive Aspects of the Prototype. In addition to the points outlined above, participants highlighted a number of positive aspects emerging from the prototype: (i) the tool was commended for its **overall user-friendliness**, including its clear layout and ease of use (e.g. navigating through the video clips); (ii) the **presentation of data** was highlighted as a positive point, i.e. the fact that everything that is needed to create content descriptions was presented on one screen, obviating the need to search for metadata and enabling the user to choose and decide whether to use a given (machinegenerated) description; (iii) the **timeline** was considered to be helpful, which is interesting to note in light of the mixed scores in the features and functions section of the questionnaire; (iv) the use of different **data lanes (tiers)** to display the various types of metadata was thought to be useful (this is corroborated by the ratings given in the features and functions section of the questionnaire); (v) the ability to **type a description while the video clip was running** appealed to participants; (vi) the availability of **separate fields** for automatically generated data and human-made descriptions was deemed

useful for accuracy and reliability; (vii) **traceability** (i.e. the possibility to see whether the material originated from another clip or program) was highlighted as important in the context of re-use rights, with participants pointing out that it can be problematic if re-use rights are not clear for a clip that is re-used or re-sold.

Aspects of the Prototype Requiring Improvement. Conversely, there were aspects of the prototype which participants considered could benefit from further improvements: (i) participants in one focus group felt that the platform should offer **more room to tell a story**, e.g. by incorporating a storyboard function; (ii) other participants queried the helpfulness of entering free keywords, suggesting that lists of **controlled vocabulary** would be more effective; (iii) another concern was that **repeated descriptions** were required across contiguous segments meaning that some labels had to be typed repeatedly, suggesting that a copy and paste functionality would be important; (iv) there was some uncertainty as to whether a description, tag or session had been **saved**, which led to re-entering descriptions repeatedly and (v) as was apparent from the specific features and functions section of the questionnaire, the **'set in set out' time feature** was difficult to use.

Principal Functions of the Prototype. Overall, all existing functionalities were considered useful but the detailed assessment and comments on some of the features varied in line with the practices and requirements of the participants' companies. This section summarises the main points made by the participants.

Content Descriptions. Asked how they would like to enter the content descriptions, i.e. whether they would prefer to edit/overwrite suggested descriptions or to write the descriptions from scratch, the participants said that this depends on the quality of the automated captions, as it would take time to edit highly incorrect descriptions. Some participants wondered whether this decision may also depend on the type of the content being described. A general view was that it would be useful and save time to have **suggestions** (of reasonable quality), as long as they can be easily and quickly deleted, if necessary. Paraphrased comments from participants also included the following, although they should be interpreted with caution, as they may have been influenced by the quality of the video captions in the sample material: (a) full sentences seem to be difficult to produce automatically; it would therefore be more helpful if the machine gave **keywords** which could be deleted or confirmed, as descriptive sentences are written; (b) writing content descriptions as **full sentences** meant it can take longer to correct the automated data than to write descriptions from scratch without automated prompts; (c) automatic suggestions are not always useful for content description, but **tags** may be useful to the end users; (d) for some users, since editing required both **deletion** and at other times **activation**, this increased the complexity of processing data (other participants, by contrast, regarded the way in which suggestions are now available as interesting and easy to use).

Segmentation and Shot Changes. Participants who describe shots or similarly short segments in their normal work practice found the video segmentation helpful and used it in the evaluation, while others who do not use it on a daily basis found it too detailed (the minimum length of a segment for description is normally five seconds). The segmentation feature was seen as particularly useful for programmes with several sub-topics, as it helped users move through the clip in an ordered fashion.

Transcription and Translation. Generally, participants found this feature useful as long as the transcript was accurate. Participants who do not normally describe speech said they would not need this feature in their own practice, but they could imagine cases where it could be useful, particularly for talk shows, factual programmes and noteworthy quotes. Transcription also

enables the describer to **check names of persons and places and quotes**, and assists with (voiceover) **script writing**. Likewise, it would be very helpful for subtitling for the deaf and hard of hearing. Translations were found helpful for the description of **foreign-language programmes**, as a way of gisting. Interestingly, participants who describe the visual content only, without referring to any speech in the video clips, pointed out that they still need to understand what is said in the video clip, as context for the descriptions of the visuals. The translation feature was deemed to be particularly useful for broadcasters with multilingual programmes and archives.

Face Recognition. Participants found the idea of automatic face recognition [8] useful especially for describing **old material without metadata**, but questioned its reliability. Another use that was highlighted was the **identification of foreign names** through face recognition, as this would help with spelling if delivered as an editable functionality.

Tags. The automated tags were described as a very useful feature for both content description and retrieval. The general view was that the tags **save time** and **prevent typos** in the description. One participant also explained that the tag field enabled him/her to enter individual key words which would have been too fragmented in the description fields. S/he felt that tagging and content description supported each other. As with other automatically generated data, the problematic quality of some tags was highlighted (e.g. wrong place names and geodata lacking precision).

Additional suggestions for improvement made by participants. Finally, the participants also suggested functions and metadata that could be added into the tool (Table 7):

Table 7. Suggestions for additional functionalities and metadata

Feature	Notes
Storyboarding	For an 'at a glance' overview of segments
Topic capture	Speech alone may not contain a topic
Vocabulary or ontology	To assist with description creation
Supplementary speaker notes	E.g. to note when a speaker is not seen
Supplementary shot data	E.g. internal/external; type of shot/ratio
Automatic music recognition	To assist with rights clearances
Automatic building identification	
Extra time lane/tier for subtitles	
Author field	
Confidence estimation for machine data	Automated caption reliability indicator
Provenance of data	Capacity to note origins of film material

4 Discussion

4.1 Considerations for Future User-Interface and User Experience Improvements

Clearly, this evaluation was primarily aimed at the *Flow* prototype, i.e. to explore the extent the prototype's ***functionality*** can support the production of AV content descriptions based on machine-generated metadata and video captions, human creation and human post-editing. As such, the evaluation reveals an overall very positive perception of the tool by professional content describers with different levels of experience and from different company backgrounds. Based on their own experience with similar platforms, the study participants engaged positively with the tool, highlighted benefits and made a number of suggestions for the further improvement of existing functionality and integration of additional functionality. In addition to the direct outcomes of the study, the high level of engagement can also be taken as

a positive sign, which is in line with the changes currently taking place in the participants' work environments, including changes to archival systems, which in their view make the development of the *Flow* platform highly timely.

Beyond a functional evaluation of the platform's features, the study was also aimed at understanding the extent to which this novel **workflow** constitutes a viable way of producing AV content descriptions in the context of archive retrieval and re-sale. In this respect, the main finding emerging from the evaluation is that professional content describers acknowledge the benefits of this workflow as long as the machine-generated content is of a sufficient quality.

However, the evaluation also reveals broader conceptual issues about this workflow. One particularly interesting point is the perception by some of the study participants that a content editing platform should enable them to **"tell a story"** and, consequently, that a storyboard function would be helpful in achieving this aim, i.e. in gaining an overview of the entire video clip they are describing. This tallies with our earlier research [1, 2] on explicating the human process of discourse comprehension and production, which demonstrated that the process is holistic, requiring continuous attempts at creating a mental representation of the story emerging from any given (verbal or multimodal) text. As such, we use cues from this text to activate common knowledge which, in turn, helps us to integrate all elements present in a discourse to form a coherent storyline. Familiar with describing video footage at this more abstract, integrated level, some describers took issue with the **amount of detail** the automated captions provided and with their fragmented, disjointed nature. This corresponds to our observation that video captions currently offer only basic descriptions of the visual content, as opposed to offering an event narration [1, 2]. Some participants did, however, point out that the creation of descriptions for archive retrieval/re-sale purposes requires them to focus on individual physical objects in the video footage.

In the short- to mid-term, one of the main benefits of the novel workflow could be that it helps a human describer deliver the required physical descriptions consistently and efficiently (provided that accuracy of object recognition can be further improved), whilst also supporting the human describer in **contextualising** and **interpreting** the material without which even the most accurate object recognition will fail to be meaningful. A useful next step for developing the editing platform may therefore be to provide support for the contextualisation work (that is, for "storytelling"), in the form of storyboarding tools or similar aids to holistic meaning-making.

Another implication from the participants' comments is that the **level of detail** included in any description and the decision about what constitutes the most useful combination of automation and human work is governed by the purpose of the description (and the audience) as well as the type of material (e.g. contemporary vs. legacy material). The immediate conclusion from this would appear to be that the editing platform would be most useful if it offered a high level of **customisation** to cater for different purposes and settings. Customisable options would also cater for **personalisation** to accommodate different working practices and styles, as well as the user's own preferences. Ultimately, customisation options in the tool should also pave the way for a future version of the editing platform that can cater for both the more object-oriented type of description mostly required for archival purposes, and the more narrative approaches required for other types of AV content description that this project has considered, namely audio description for visually/cognitively impaired audiences, which in addition to the differences in description style, also comes with a raft of technical requirements (e.g. fitting the description in silent moments in the audio track).

4.2 Limitations of the Study

Discrepancies in the data. Triangulation of the different sets of data suggests that there are some **discrepancies** between what participants reported in the FG and questionnaire, and what they actually did during the hands-on session. In some cases, participants seem to have misremembered or were not entirely clear about what they did when they worked with the prototype (for example, issues with saving content descriptions). Such problems were to be expected, given that the participants only learned how to use the prototype during the pre-study workshop.

Confounding Factors. Whilst the evaluation was focused on the usability and functionality of the prototype, many participants raised related topics such as the issue of **accuracy** and **reliability** of the automatically generated metadata and content descriptions. It is possible that the quality of the sample video captions and person/location [9, 10] content affected participants' perceptions of the prototype's usability and functions to some extent. The mismatch between the **language** in which the automatically generated video captions were presented (English) and the participants' native languages, may have been a further confound. Finally, as the study took place as virtual sessions, with participants working from home due to the Covid-19 pandemic, some reported problems with their technical set-up at home. This may have flavoured their opinions of the platform's useability.

5 Conclusions

The main aim of this evaluation was to find out to what extent, in the perception of professional content describers, the *Flow* prototype supports the creation of **viable video descriptions** and the way individual workflow and functionalities of the platform are perceived by professional describers. This was viewed as a preliminary step towards a secondary 'end-goal' of creating semi(automated) audio description which, it was acknowledged, is a considerably more complex task. Through a mixed method approach we have been able to measure the perceptions of the participants in relation to the key dimensions of **user experience** and **usability**, and **essential features** of the prototype, as well as capturing qualitative feedback on the **human-machine workflow** including suggestions for improvement. The prototype received an **overall positive evaluation** and was found to be capable of **supporting the task of describing AV content** for the specified purpose. As a **tool**, the prototype was found easy to use, intuitive, functional and logical. Regarding the validity of the **workflow** that the prototype has been developed to support, participants acknowledged that this was a useful way forward, i.e. a novel technology-enhanced human workflow. Looking to the future, **data quality** will be the primary driver of progress in this area, with automated captions playing a useful role in identifying aspects of the content that a human operator might overlook, as well as in improving consistency of description and bringing legacy material lacking metadata into the fold of digitally searchable media archives. It is also an early first step in developing a methodology for delivering **semi(automated) audio descriptions**. Whether such workflows would save time and improve the quality of the descriptions was not measured in this evaluation, due to the low quality of the automated data. Nevertheless, the present evaluation is the first of its kind (to the researchers' knowledge) – i.e. an evaluation of a prototype tool that supports post-editing of automatically generated textual content (including metadata and narrative), produced in an automated process of *intermodal* translation (images to text).

6 References

1. Braun, S., Starr, K.: Finding the right words: investigating machine-generated video description quality using a human-derived corpus-based approach. J. Audiov. Transl. **2**(2), 11–25 (2019). https://doi.org/10.47476/jat.v2i2.103
2. Starr, K., Braun, S., Delfani, J.: Taking a cue from the human: linguistic and visual prompts for the automatic sequencing of multimodal narrative. J. Audiov. Transl. **3**(2), 140–169 (2020). https://doi.org/10.47476/jat.v3i2.2020.138
3. Huang, T.H., et al.: Visual storytelling. In: Proceedings of NAACL-HLT, San Diego, California, 12–17 June, pp. 1233–1239 (2016). https://doi.org/10.18653/v1/N16-1147
4. Park, J.S., Rohrbach, M., Darrell, T., Rohrbach, A.: Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6598–6608 (2019). https://doi.org/10.1109/CVPR.2019.00676
5. Laaksonen, J., Guo, Z.: PicSOM experiments in TRECVID 2020. In: TRECVID 2020 Workshop, 17–19 November, Online Conference (2020)
6. Limecraft homepage. https://www.limecraft.com/. Accessed 09 June 2021
7. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89350-9_6
8. Lisena, P., Laaksonen, J., Troncy, R.: FaceRec: an interactive framework for face recognition in video archives. In: 2nd International Workshop on Data-driven Personalisation of Television (DataTV) Collocated with the ACM International Conference on Interactive Media Experiences (IMX 2021), 21–23 June 2021, forthcoming. https://doi.org/10.5281/zenodo.4764633
9. Harrando, I., Troncy, R.: Named entity recognition as graph classification. In: Verborgh, R., et al. (eds.) ESWC 2021. LNCS, vol. 12739, pp. 103–108. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80418-3_19
10. Porjazovski, D., Leinonen, J., Kurimo, M.: Named entity recognition for spoken finnish. In: Proceedings of 2nd International Workshop on AI for Smart TV Content Production Access and Delivery (AI4TV), pp. 25–29 (2020). https://doi.org/10.1145/3422839.3423066