memad.eu
info@memad.eu

Twitter - @memadproject
Linkedin - MeMAD Project

MeMAD Deliverable

D6.9 – Evaluation report, final version

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	23.6.2020
Start date of the project	1.1.2018
Due date of the deliverable	31.03.2021
Actual date of submission	28.04.2021
Lead beneficiary for the deliverable	Limecraft
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
Karel Braeckman	Limecraft	karel.braeckman@ limecraft.com
Jaleh Delfani	University of Surrey	j.delfani@surrey.ac.uk
Maija Hirvonen	University of Helsinki / Tampere University	maija.hirvonen@tuni.fi
Maarit Koponen	University of Helsinki	maarit.koponen@helsinki.fi
Sacha Lagrillière	YLE	sacha.lagrilliere@yle.fi
Lauri Saarikoski	YLE	lauri.saarikoski@yle.fi
Kim Starr	University of Surrey	k.starr@surrey.ac.uk
Umut Sulubacak	University of Helsinki	umut.sulubacak@helsinki.fi
Liisa Tiittula	University of Helsinki	liisa.tiittula@helsinki.fi
Tiina Tuominen	YLE	tiina.tuominen@yle.fi
Kaisa Vitikainen	YLE	kaisa.vitikainen@yle.fi
Dieter Van Rijsselbergen	Limecraft	dieter.vanrijsselbergen@ limecraft.com

Document reviewers		
Name	Beneficiary	e-mail
Mikko Kurimo	Aalto University	mikko.kurimo@aalto.fi
Jörg Tiedemann	University of Helsinki	jorg.tiedemann@helsinki.fi

Document revisions			
Version	Date	Authors	Changes
0.1	30/12/2020	Kaisa Vitikainen, Maarit Koponen	Draft of general evaluation methodology and setup for subtitling evaluations.
0.2	09/02/2021	Kaisa Vitikainen, Maarit Koponen, Tiina Tuominen, Umut Sulubacak	Final draft of Section 8.1, 8.2 and 8.3 on the subtitle evaluations.
0.3	28/03/2021	Sabine Braun, Kim Starr, Jaleh Delfani, Liisa Tiittula	Final draft of Section 9 on the content description evaluation.

0.4	03/04/2021	Sacha Lagrillière, Maarit Koponen, Umut Sulubacak, Dieter Van Rijsselbergen	Final draft of Sections 6 and 7 on video editing and content retrieval evaluations.
0.5	05/04/2021	Karel Baeckman	Input regarding potential improvements to content description editor application.
0.6	08/04/2021	Maarit Koponen, Umut Sulubacak	Clarifications on pipelines and evaluation setup for subtitling evaluations.
0.8	19/04/2021	Dieter Van Rijsselbergen	Completed draft, ready for review.
0.9	26/04/2021	Dieter Van Rijsselbergen	Final text, with fixes from reviews.
1.0	27/04/2021	Dieter Van Rijsselbergen	Final document, with appendices and final revisions on Section 10.

Abstract

This is the third and final evaluation report of the MeMAD prototype integrated platform and it reports end user feedback on the prototype and the underlying components that it makes use of. In this round of evaluation, we evaluated implementations of searching and browsing for content in ingested and archived content, video editing assistance using multi-modal and multi-lingual metadata, and auto-generation of intralingual and interlingual subtitles for a second time. Additionally, we describe the results of two new evaluations, namely on the consumer reception of automatically generated interlingual subtitles and on the usability of our prototype editing application for human-in-the-loop content description.

We describe in detail how each evaluation study was conducted and discuss the user feedback and data collected for each study. Additionally, we deduce overall conclusions on each of the evaluated functional epics implemented in the MeMAD prototype, and we make a final assessment of the usability and maturity of each content enrichment process employed in the MeMAD project.

The main outcome of this evaluation round remains a positive result in that the test panels positively evaluated the potential of the implemented services, especially in content retrieval, content description authoring and intralingual subtitling, but that work remains to be done on improving a variety of hindrances to fully exploit the possible gain in usability and efficiency of the services developed in the project.

Contents

1 Introduction

In this deliverable, the third of three report iterations, we describe the results of the last set of evaluation rounds for the prototype MeMAD platform as described by the complement of D6.5 and D6.8, and performed as part of task T6.3 in WP6. This deliverable reports on the user feedback and observations gathered by the consortium during the user panel evaluations that took place during the final fifteen months of the project.

The discussed round of testing evaluated all components that were delivered in a mature form in the run-up to the end of the project year, either in an improved form from the previous evaluations, or in a first completed version. It comprised of the following functional evaluations:

1. Searching and browsing for content in ingested and archived content, of which the aim was to learn to what extent metadata automatically generated by the various MeMAD feature extraction tools can serve end users in locating content, and how they augment or replace existing metadata in finding materials from an archive or a production system. We collect these tools under the general denomination of **Automated Metadata Extraction** or **AME** tools;
2. Editing assistance using multi-modal and multi-lingual metadata, of which the aim was to learn how video editors are helped by the aforementioned AME-generated metadata during the video editing process;
3. Auto-generation of subtitles, which we evaluated to learn to what extent – and how convenient and intuitive – the manual correction of automatically generated subtitles can improve the efficiency of the subtitling process, both in the context of intralingual subtitling (in which the subtitles are authored in the same language as the spoken language) and interlingual subtitling (in which the created subtitles serve the audience in language different from that of the original audiovisual content). In this evaluation cycle, we also evaluated the audience's perception of these auto-generated subtitles, with a focus of taking the human author out of the loop. Questions we answered were how acceptable raw machine-translated and fully automated subtitles are to viewers and which are the key factors that determine subtitle quality and acceptability from viewers' point of view. Also, we measured how much cognitive load raw MT-translated subtitles cause for viewers compared to professionally produced subtitles;
4. An evaluation of the prototype content description application developed in T5.4, as a joint effort of WP5, WP6 and indirectly WP2 for supporting AME services. The end user evaluation followed a first evaluation of the application's design (as described in D5.4) and was then trialed with users experienced in content annotation and archival processes. The prototype aims to support a workflow in which the human remains a crucial element in the loop, with the curation of functional video descriptions generated first by AME services and then corrected manually. Given that, we sought to answer questions such as how the prototype supports the human-in-the-loop workflow and which of the tool's features are perceived to be most/least beneficial in this process.

To provide the reader with additional background, we additionally summarize how this deliverable compares to the previous evaluation iteration report (i.e., D6.6), in Section 2. Then, in Section 3, we repeat the context regarding the place of the evaluations in the overall human-centered design and implementation process adopted for the development of the MeMAD prototype. Section 4 explains precisely which functionality implemented by the MeMAD prototype was evaluated and how that relates to the functional requirements laid out beforehand for the prototype's services (cf. also D6.4 and D6.7), thereby providing context to the overall executed project work plan.

Section 5 describes the general evaluation methodologies that we adopted for performing the end user evaluations. In Sections 6 through 9 each functional evaluation is discussed. Each section follows a similar structure: we introduce the evaluation and motivate why it was undertaken, we discuss the user test setup and the executed evaluations tasks, and then we provide an analysis of the collected data, split among each type of data collected (questionnaires, interviews, usage metrics, etc.).

In Section 10 we take a broader look at the evaluation results in order to express our overall findings of the project's prototype and its underlying technologies. We conclude this deliverable in Section 11.

2 Changes with regards to deliverable D6.6

This deliverable is a report that describes the final round of evaluations of the matching completed final MeMAD prototype platform version, of which the first iteration was documented in deliverable D6.3 and the second one in D6.6. With respect to the D6.6 report, the following changes have been made:

1. Section 4 was updated to list how the evaluations undertaken match with the finished functionality of the prototype platform and how that functionality has been evaluated;
2. Section 5 updates the process methodology used for this round of evaluations, which required adaptation for three reasons: due to findings from the previous evaluations, due to added functionality evaluated and unfortunately also as a result of the COVID-19 pandemic situation which forced the adoption of different user interactions from those employed before.
3. The contents of Sections 6 through 9 has been replaced with a description of the newly undertaken evaluations.
4. Due to this being the final evaluation report of the project, the scope of the impact section 10 has been revised; this deliverable now provides a general conclusion on each implemented functional epic, on each underlying automated metadata extraction technology (AME) and on the impact of the new state of the art on future media production processes.

3 Methodology for the MeMAD prototype platform development

This section describes the methodology followed in the execution of the work in Work Package 6, and as such also for the evaluation of the second MeMAD prototype, which forms an important part of validating the designs and implemented software components. While the overall methodology has already been explained in D6.1, we reiterate those sections relevant to the work described in this deliverable. The MeMAD development and evaluation methodology is built on two pillars:

1. As a guiding principle for implementing the functionality of the project's prototype and its underlying individual components we use the four project use cases (PUCs) and their derivative user stories and development epics defined in the project's Description of Action (DoA) and in D6.4 and D6.7.
2. With regard to the User-centered Design (UCD) methodology[1] that we adopted, the evaluation of the prototype is part of the 3rd phase. The execution of this methodology in MeMAD occurs in several steps, as illustrated by Figure 1.
    - In Phase 1, the context of use across the entirety of media production and consumption process was investigated and subsequent actual functional requirements were defined. More detailed user requirements have been described as user stories to explain more specific sets of desired functionalities, each of them fitting within the definition of one of the Project Use Cases (cf. D6.4 and D6.7).
    - As part of Phase 2, based on the finalized list of relevant user stories, the exact requirements involved for each story were further refined. The results from this effort are both functional and non-functional requirements which serve as the basis for coordinating the development efforts in T6.2. Deliverables D6.1, D6.4 (in intermediary form) and D6.7 (in final form) provide these guidelines.
    - The actual work in T6.2, the development of the prototype also falls under the 2nd phase of the UCD process, which has now been completed through its last cycle, based on the second round of evaluations reported in D6.3 and D6.6 and the subsequent revised specifications from D6.7. The efforts performed under T6.2 in the period from M24 to M39 are the subject of deliverable D6.8. The selection of D6.7 functionalities for the final iteration was done at the consortium meeting at INA in February 2020. Final confirmation and implementation choices were made at subsequent online consortium meetings organized in June 2020 and October 2020.
    - Finally, in Phase 3, each prototype platform iteration has been evaluated with end users, first to verify the proper implementation of the (non)functional requirements, and secondly to provide improvement feedback to the design process such that a final development cycle can be implemented for the final prototype version.
      Multiple cycles of Phase 2 implementations and Phase 3 evaluations have been completed in the project. The evaluation work was performed in Task T6.3, and this deliverable is the report of the last Phase 3 evaluation cycle.

[1] Cf. ISO Standard 9241-210:2010 – Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems.

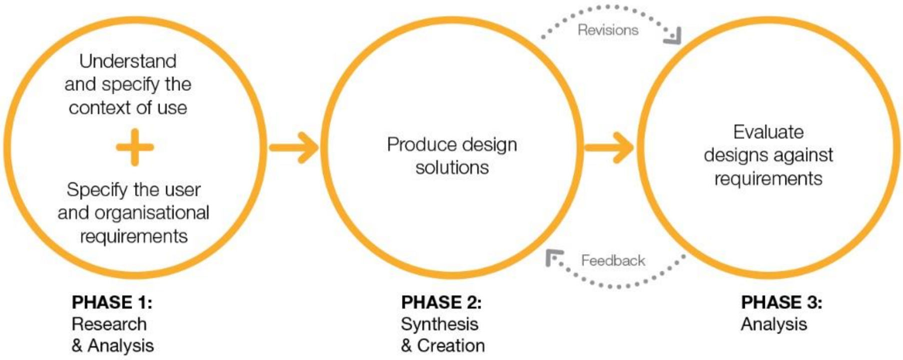

Revisions
Understand and specify the context of use
+
Specify the user and organisational requirements
Produce design solutions
Evaluate designs against requirements
Feedback
PHASE 1:
Research & Analysis
PHASE 2:
Synthesis & Creation
PHASE 3:
Analysis

Figure 1: Synopsis of the User-Centered design (OCD) process (from O'Grady, 2008[2]).

As explained above, because pieces of the UCD process encompasses the entire duration of the project, this deliverable forms the final part of the 'puzzle' for this work package and rounds off T6.3 with the final conclusions concerning the MeMAD prototype.
We summarize where each piece of work was oriented in which deliverable of Work Package 6 in Table 1.

[2] Cf. Visocky O'Grady, J. & Visocky O'Grady, K. (2008) The information design handbook. Mies: RotoVision.

Interchange format specification and requirements definition	**The MeMAD prototype**	**Evaluation of the MeMAD prototype**
D6.1: Definition of the context of use and an initial set of high-level user requirements. In addition, this deliverable maps out a first revision of required metadata and sets the requirements for the first prototype iteration (M3).	**D6.2:** A report on the first implementation of the prototype, executed per the specifications of D6.1 (M12).	**D6.3:** An evaluation of the first prototype and its requirements, to the extent possible with the limited implementation. This report also includes feedback concerning the use cases and requirements for exchange format specifications (M12).
D6.4: Refinements of the initial set of high-level user requirements based on feedback from external advisors. This second version will define more detailed requirements for the second MeMAD prototype, including test criteria and scenarios (M18).	**D6.5:** A report on the implementation of the second prototype, executed per the specifications of D6.4 (M24).	**D6.6:** An evaluation of the second prototype and its requirements (M24).
D6.7: Definition of the final requirements and test criteria for the MeMAD project prototype, along with final specifications of all metadata exchange formats (M27).	**D6.8:** A report on the implementation of the final MeMAD prototype, executed per the specifications of D6.7 (End of Project).	**D6.9:** A report on the evaluation of the final MeMAD prototype, which will be done by both the consortium and interested parties outside the project consortium (End of Project).

Table 1: Orientation of MeMAD Work Package 6 deliverables.

4 Use cases functionality evaluated in the second MeMAD prototype platform

In this section, we summarize the MeMAD prototype functionality that was evaluated in the final evaluation cycle of the project. This iteration was guided by the requirements defined in D6.4 and D6.7 and was focused on two pillars, namely refining the previous evaluations of end-user functionalities and executing evaluations for newly developed functionalities in the final project prototype version. While the implemented functionality of the platform is extensively described in D6.8 (cf. highlighted boxes in Figure 2), we summarize these functionalities here as a context for describing the various evaluation tasks that were executed.

With the final iteration of the prototype platform and its components available, the evaluation activities undertaken during the final project period, namely from M24 – M39, have been summarized in subsections 4.1 through 4.7. An extensive discussion of each evaluation is then given in later sections of this deliverable.

In short, we can summarize that the functionality of the final MeMAD platform iteration has been evaluated in the following four groups of evaluations:

1. A 2nd round of evaluations for Epic 6.3 (“Searching and browsing for content in ingested and archived content”), in particular user stories 2.2.1, 2.2.2 and 2.2.5, which is covered in Section 7;
2. A 2nd round of evaluations for Epic 6.5 (“Editing assistance using multi-modal and multi-lingual metadata”), in particular user stories 2.1.5 and 2.1.6, which is covered in Section 6;
3. A 2nd round of evaluations for Epic 6.11 (“Auto-translation of subtitles”), in particular user stories 4.1.4, 4.3.1, 4.3.2, 4.3.3 and 4.3.4, which is covered in Section 8 and its three subsections. This evaluation was split across evaluations for the production of subtitles (from the content producer’s or subtitler’s point of view, cf. Section 8.1 and 8.2), and for the reception of subtitles (from the audience’s point of view, cf. Section 8.3);
4. A 1st round of evaluations of the fully implemented and functional prototype application developed for Epic 6.10 (“Auto-generation and correction of content descriptions”) following a first evaluation of the application’s design (as described in D5.4). This concerns in particular user stories 4.2.2 and 2.2.4, and is covered in detail in Section 9.

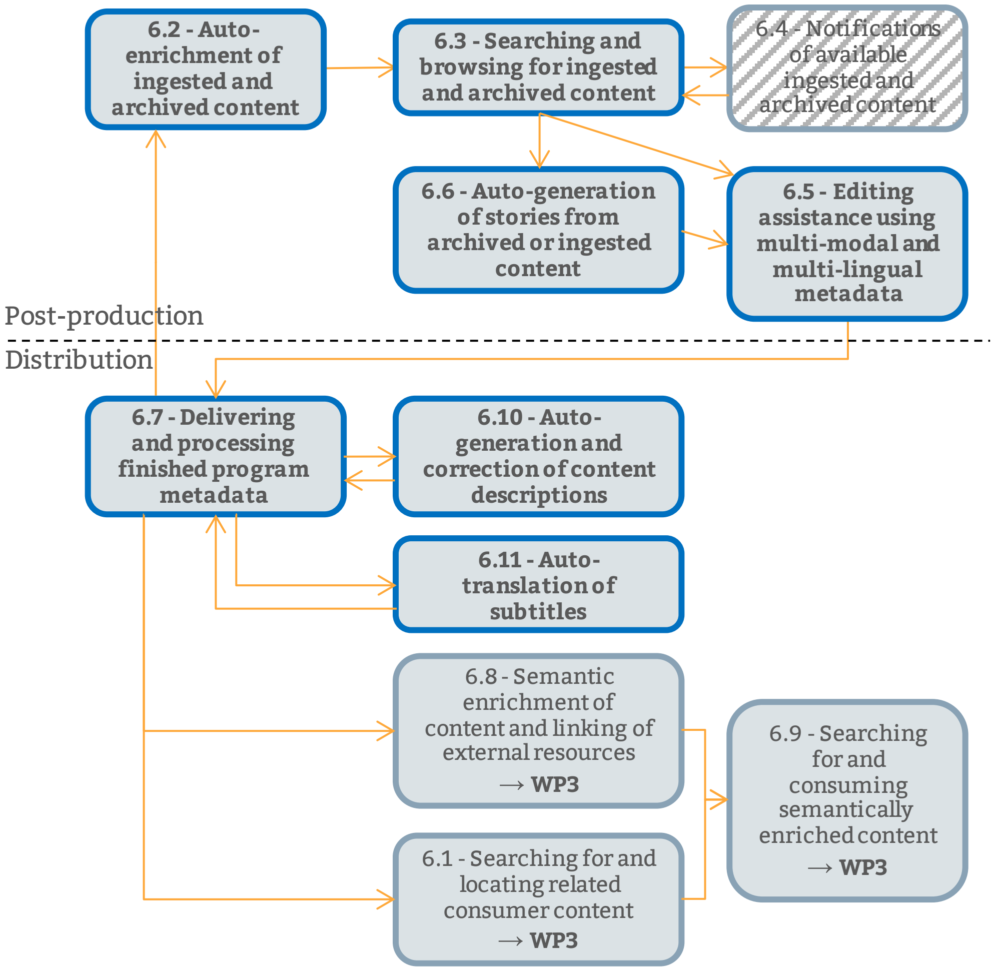

6.2 - Auto-enrichment of ingested and archived content
6.3 - Searching and browsing for ingested and archived content
6.4 - Notifications of available ingested and archived content
6.6 - Auto-generation of stories from archived or ingested content
6.5 - Editing assistance using multi-modal and multi-lingual metadata
Post-production
Distribution
6.7 - Delivering and processing finished program metadata
6.10 - Auto-generation and correction of content descriptions
6.11 - Auto-translation of subtitles
6.8 - Semantic enrichment of content and linking of external resources → WP3
6.9 - Searching for and consuming semantically enriched content → WP3
6.1 - Searching for and locating related consumer content → WP3

Figure 2: Functional epics implemented in the final MeMAD prototype platform.

4.1 Functionality and evaluations for Epic 6.2

Table 2 lists the user stories implemented in the MeMAD prototype for *Epic 6.2 – Auto-enrichment of ingested and archived content*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	Implementation	Evaluation
2.1.1 - Real-time analysis and indexing of ingested content, 2.1.2 - Extensive analysis of ingested content.	Added and improved integrations of backend services provided by the consortium into the prototype allow automated (and in many cases real-time) processing and subsequent enrichment of audiovisual content that has been ingested into the platform. For this final iteration, the focus of integrations was on services that deal with the visual content analysis modality (e.g., OCR, face recognition and scene classification),	The outcome of the processing components is used by end users for searches or as a starting point for authoring subtitles or content descriptions. As such, this functionality was always evaluated in the context of a functional end user process, incl. the evaluation of Epics 6.3 and

	and multi-modal enrichment (in particular 'Deep Captioning' of content and language identification). Some existing integrations were improved and extended. Overall, all envisioned and required services have been integrated and contribute to a broad spectrum of useful content enrichment onto which other functional epics are constructed.	6.5 and indirectly also for Epics 6.6, 6.10 and 6.11.
2.2.4 - Intuitive manual correction of automatically generated metadata.	Intuitive user interfaces were developed for post-editing audio transcripts and face recognition metadata. Further, a large effort was made in collaboration between WP6 and WP5 to build a dedicated content description editor prototype application for easier manipulation of various metadata. This topic is discussed in depth in D5.4 and D6.8.	Idem.

Table 2: Implemented functionality and evaluations of Epic 6.2.

4.2 Functionality and evaluations for Epic 6.3

Table 3 lists the user stories implemented in the MeMAD prototype for *Epic 6.3 – Searching and browsing for ingested and archived content*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	**Implementation**	**Evaluation**
2.2.1 - Searching for content in archives.	We have built support for this use case, centered around locating content in an archive through search actions, using various metadata that were added to the prototype system either with human-curated descriptions, or through automatically generated enrichments. The search actions are supported by an extensive search indexing and querying system and various user interface elements to help users browse search results and act on these results to submit them to down-stream production processes.	Searching and browsing for content and parts of content in audiovisual archives has been evaluated in a second round by end users using the prototype's user interface, fueled by a combination of both legacy and expanded sets of auto-generated metadata. This extended metadata was produced by additional AME services integrated into the final MeMAD prototype platform.
2.2.2 - Searching for segments of content in archives.	As an extension to 2.2.1, the platform also supports searching and browsing of temporally segmented content based on similarly segmented metadata (e.g., audio transcripts and face detections). In the final prototype, more temporally segmented metadata such as OCR text recognition and scene classification have	

	been added for more diverse search results. Selections and exports can also be made on a subclip level basis.	This evaluation track is discussed in Section 7.
2.2.5 - When looking up archival content, hyperlinked related media are also shown.	As part of the named entity disambiguation process, links to relevant knowledge bases (such as Wikidata[3] or DBpedia[4]) are added to the metadata such that users can immediately look up concepts and further information using these links. D6.5 provides more information on this topic.	

Table 3: Implemented functionality and evaluations of Epic 6.3.

4.3 Functionality and evaluations for Epic 6.5

Table 4 lists the user stories implemented in the MeMAD prototype for *Epic 6.5 – Editing assistance using multi-modal and multi-lingual metadata*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	Implementation	Evaluation
2.1.5 - Editing assistance using multi-model metadata.	Various metadata from the platform, including audio transcripts, named entity metadata, face detections, etc. are exported in a relevant format to bring them into a professional editing environment where they assist the editor in her editing tasks.	A second round of evaluations were executed to re-test the usability of providing auto-generated metadata in dedicated video editing environments, commanded by professional video editors. Starting from an improved export mechanism, we evaluated whether this new setup improves working with this metadata in a professional craft editing environment. This evaluation is discussed in Section 6.
2.1.6 – Use of auto-translated content for editing.	Same-language and machine-translated audio transcripts are exported and presented to editors such that they can make use of the translation to give them insights into the selected audiovisual materials.	

Table 4: Implemented functionality and evaluations of Epic 6.5.

[3] Cf. Wikidata: “a free and open knowledge base that can be read and edited by both humans and machines”, available at: https://www.wikidata.org/wiki/Wikidata:Main_Page.

[4] Cf. DBpedia, available at: https://wiki.dbpedia.org/.

4.4 Functionality and evaluations for Epic 6.6

Table 5 lists the user stories implemented in the MeMAD prototype for *Epic 6.6 – Auto-generation of stories from archived or ingested content*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	Implementation	Evaluation
2.2.6 - Users are presented with auto-summarized segments of archival content.	For this user story, the final iteration of the platform prototype has been extended with a pipeline and custom workflow for automated content segmentation. Pieces of content that deal with a single topic and that are visually similar, are grouped to form the basis for further automated summarization.	The implementation done for this epic is still quite experimental and needs more exploration before a wide-scale evaluation of the pipeline's performance can be organized. The content segmentation functionality was trialed in a proof-of-concept with YLE in the final project months. Findings from this PoC are discussed in detail in D7.4. Performance of other algorithms involved, e.g., Media Memorability are discussed in D3.3. Unfortunately, the overall feature set for story building realized at the end of the project was deemed to immature to organize an extensive evaluation for. More future exploration and testing of individual components is required before a fully integrated workflow can be trialed.
2.2.7 - Users are suggested auto-generated stories from archive content that they can modify and shape into final program items.	The framework built for auto-segmentation of content provides the foundations for further enrichments, such as media memorability. Using these enrichment scores, the platform can group search results that match the intended topics into a story that can be further shaped and manipulated by end users, either within the user interfaces provided inside the platform (i.e., the 'story builder'), or by exporting these stories to external editing systems.	

Table 5: Implemented functionality and evaluations of Epic 6.6.

4.5 Functionality and evaluations for Epic 6.7

Table 6 lists the user stories implemented in the MeMAD prototype for *Epic 6.7 – Delivering and processing finished program metadata*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	Implementation	Evaluation
2.3.2 - Delivering relevant production	More export mechanisms were implemented to allow the MeMAD platform to export production metadata to downstream	Delivering production metadata to downstream processes has not been evaluated explicitly in WP6,

metadata downstream.	production processes to external systems. These metadata include, amongst others, shot-cut-boundary information created by the platform's media processing services, speech transcripts, and subtitles generated and manually authored within the platform, and curated and approved content description metadata.	not as far as studies of end user processes are concerned. Functionally correct exports of metadata have been tested indirectly as part of the development of various functional epics though (e.g., exchange of metadata with edit stations, exchange of subtitle files for intralingual subtitling evaluations, etc.).
2.3.3 - Processing and harmonizing delivered production metadata.	The platform has been extended with metadata imported from the legacy content metadata stored in the MeMAD Knowledge Graph (cf. D3.2). This provides the capability of importing this information from a harmonized and standards-based data set and using it as a content enrichment source to support the other user stories implemented by the MeMAD platform. This functionality is discussed further in D6.8 and in D6.5.	As with the metadata obtained through the MeMAD ingest enrichment services (cf. Epic. 6.2), the results of the Knowledge Graph integration makes legacy metadata available for searching and browsing archived content. As such, it is indirectly evaluated in Section 7.

Table 6: Implemented functionality and evaluations of Epic 6.7.

4.6 Functionality and evaluations for Epic 6.10

Table 7 lists the user stories implemented in the MeMAD prototype for *Epic 6.10 – Auto-generation and correction of content descriptions*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	**Implementation**	**Evaluation**
4.2.1 - Content consumption with auto-generated audio descriptions.	Given the decision for the development of algorithms and the application prototype to focus on the generation and curation of content descriptions for supporting content retrieval tasks (cf. D5.3 and D5.4), no further work was done in WP6 to support this particular use case.	No evaluation was done as this use-case was discarded.
4.2.2 - Manual corrections improve auto-generated audio descriptions.	Implementation-wise, we can consider the work on both user stories together. While not geared toward producing audio descriptions, the content description editor prototype application (cf. D5.4) does handle its own set of metadata paradigms in terms of content description and is such optimized for a specific media production process. Further development, based on improved underlying unsupervised captioning	An extensive evaluation of the content description editor prototype was conducted in a joint WP5-WP6 effort. Various usability aspects of the prototype's GUI and source AME
2.2.4 - Intuitive manual correction of automatically		

generated metadata.	algorithms, could allow optimizations in the direction of generating audio descriptions. In the meantime, the application developed allows archivists and journalists to intuitively manipulate and correct automatically generated metadata, while at the same time delivering humanly curated and verified content descriptions. Further elaboration on this implementation is divided between D6.8 and D5.4.	metadata were evaluated by a test panel of expert users. Section 9 describes this evaluation in detail.

Table 7: Implemented functionality and evaluations of Epic 6.10.

4.7 Functionality and evaluations for Epic 6.11

Table 8 lists the user stories implemented in the MeMAD prototype for *Epic 6.11 – Intra- and interlingual subtitling*, which functionality was implemented, and how these features have been evaluated.

Applicable User Stories	**Implementation**	**Evaluation**
4.1.4 - Automated same-language subtitling.	The baseline functionality of the Limecraft Flow platform that serves as the basis for the MeMAD platform already supported the automatic generation of subtitles from audio transcripts at the beginning of the MeMAD project. It was hence trivial to include as an implementation of this user story.	Intralingual subtitling outputs from the platform were again evaluated in a second round with professional subtitlers, as elaborated in Section 8.1.
4.3.1/4.3.2 - Automatically translated subtitles for foreign users/of foreign content, 4.3.3 - Translated subtitles based on translated transcripts.	Building on the existing same-language subtitling tools that implement the 4.1.4 story, additional developments were done to bring automated subtitle translation to the MeMAD platform, based on WP2 and WP4 content and metadata processing components. This implementation discussed in detail in D6.5.	As with intralingual subtitling, improvements to the underlying technologies for auto-subtitling were re-evaluated. Interlingual subtitling, encompassing both the generation and post-editing of the subtitle and subsequent qualitative evaluations are discussed in Section 8.2. In addition, a new evaluation was conducted to test the audience's reception of automatically generated intra- and interlingual subtitles. This evaluation is discussed in Section 8.3.
4.3.4 - Manual correction of auto-translated subtitles.	Correcting and editing automatically generated translated subtitles was added to the platform to enable subtitlers to improve the quality of delivered subtitles. This MeMAD platform iteration introduces many user interface elements to perform this task, which we also elaborate upon in D6.5.	

Table 8: Implemented functionality and evaluations of Epic 6.11.

5 General evaluation setup and methodology

This section describes the common approaches and methods utilized for the evaluation of four MeMAD epics discussed in this report. The aim of the evaluations is to understand the usability of the MeMAD technologies in metadata creation, and in some cases also the metadata they created, from the user perspective. Hence, the evaluations are not based on automatic metrics but on the analysis of user experiences. For the final iteration of the MeMAD prototype, we are still dealing with technologies that are new to potential users in the creative media production industry (e.g. editors, journalists, archivists and subtitlers). Therefore, we continue in this evaluation round with a “bottom-up” approach to the study of usability and set up studies which yields insight into the perceptions, attitudes and opinions of users towards the new technology. Based on the knowledge gained from the previous (second) iteration, this final iteration repeats existing evaluation tasks with improved technologies or introduces new evaluation contexts that allow for testing the technologies developed and assessing their impact to working conditions and productivity.

This final round of evaluations was continued in the same spirit as in the second evaluation round; the participants participated hands-on with the prototypes developed, and test situations were setup so that they would be as close to authentic production situations as possible. Subtitling and editing tasks were performed with the export user software normally used by the participants in their daily work (except of course for those tasks where the user of the platform’s interface was an integral part of the task’s execution). When looking at the results of the evaluations it should be taken into account that the participants are media professionals with established workflows, working with the same software they use in their daily work in two of the evaluation tracks. As such, they are likely to be used to things working in a certain way. When things work differently than expected, adjustment is needed, which they may find difficult if they have deeply ingrained processes in place.

Our main methodological approach stems from usability research and applies the iterative design [1]. In the iterative design, we conduct repeated prototyping and testing of the MeMAD technologies, with adjustments and improvements, in order to develop the design. User testing helps to catch problems and provides feedback and thereby contributes to the overall development. The iterative design is able to track a multitude of usability issues, e.g. overall user satisfaction, different types of usability problems and task time. Thanks to these iterative improvements, the scope of the evaluations could also be extended gradually in this final evaluation round. Once the largest kinks were worked out of the interlingual subtitling software pipelines given feedback from the second evaluation round, the subtitling evaluation could now be extended with an evaluation by audiences to assess subtitle reception.

In this final iteration of the evaluation, we repeat a combination of qualitative and more quantitative approaches and gather data from users performing controlled tasks with the following methods: process data (task times) during the evaluation tasks, user questionnaires, and semi-structured interviews or focus group discussions after the tasks were completed. Some of these methods produce data on subjective evaluations by the users; their impressions, feelings, opinions, and attitudes, as explained in the following subsections 5.2, 5.3 and 5.5. To the extent possible, we supplemented these subjective assessment methods with objective evaluations, for example by quantifying user assessments with scores obtained from User Experience Questionnaires and task time measurements when testing the subtitling processes, as explained in subsection 5.2 and 5.4.

5.1 Impact of the COVID-19 pandemic

Because many of the project's evaluations were planned in the final project year 2020, and these evaluations involve users both from organizations within and outside the consortium, it is obvious that the original schedule and parts of the methodologies used were impacted by the COVID-19 pandemic. From March 2020 until the end of the project (March 2021), we had to mitigate our evaluation plans to cope with the situation that often, we could no longer evaluate with users on-site and in close proximity, and that people's working schedules were disturbed by various restrictions imposed throughout Europe over the past year. Mitigating the COVID-19 impact was one of the main reasons the project was extended by three months. Additionally, the following changes were made to the original evaluation methodology:

- The evaluation tasks in each case were set up so that the experimenters prepared the tasks and necessary files as far as possible. In the cases where the Limecraft Flow platform was used, the participants then accessed the tasks through the online platform. In the use cases where expert software was used (i.e., intralingual and interlingual subtitle post-editing cases and video editing), some of the final preparations (downloading files and importing them into the software, sending files to the experimenter after the tasks) had to be done by the participants themselves from their remote setup;
- Many users employed their personal computers for the evaluations, which meant that we no longer requested the use of keylogging software during the evaluation, in particular for the subtitling tasks. We resorted to user-reported time measurements in these cases. As the tasks were not performed in a controlled setting, this measurement is of course less precise. Task times should be compared with caution and considered more as suggestive of overall tendencies;
- The organization of Think-out Loud sessions could no longer be organized reliably and hence had to be skipped for the final set of evaluations;
- The recruitment of users was hampered somewhat, especially for evaluations planned in early in the evaluation calendar of 2020. This meant that in particular, the evaluation of Epic 6.3 was performed with less users than initially planned;
- Where possible, and in particular for the evaluation of the content description editing prototype (i.e., Epic 6.10, cf. Section 9), we used screen recording software wherever possible to provide an accurate view of how the evaluation transpired. The test panel was well informed of this data acquisition and each person consented to this recording;
- Post-evaluation semi-structured interviews and focus groups were all organized remotely using teleconferencing software such as Google Meet or Zoom.

5.2 User Experience Questionnaire

For all four use cases in this final evaluation round, an online form was used to collect subjective evaluations of the usability of the platform and outputs. The questionnaire was based on the User Experience Questionnaire (UEQ) [2]. The UEQ has been designed and widely used to elicit users' impressions, feelings and attitudes towards interactive software products. It consists of 7-point scalar evaluations of different adjective pairs (e.g. practical - impractical) describing the experience of using a product with a mid-point for neutral answers and variable labels, intended to measure both classic usability aspects and user experience aspects. For analysis and presentation of results, the analysis spreadsheets available on the UEQ website[5] were used to convert the 7-point scales into scores between -3 and +3. According to the UEQ documentation, average scores between -0.8 and +0.8 are considered neutral assessments, and values crossing these thresholds are negative or positive.

A UEQ-based questionnaire was used in the following evaluations: *Editing assistance using multi-modal and multi-lingual metadata* (Epic 6.5), *Searching and browsing for ingested and archived content* (Epic 6.3), *Intralingual subtitling – professional post-editing* and *Interlingual subtitling – professional post-editing* (Epic 6.11) and *Auto-generation and correction of content descriptions* (Epic 6.10).

Because of its direct use for the evaluation of a software application, including usability aspects, the UEQ was implemented in full for the evaluation of the *Auto-generation and correction of content descriptions* (Epic 6.10) software prototype. The specific customizations to the UEQ survey and the interpretation of its results for this study are discussed in subsection 9.1.3.

For the purposes of the other evaluations, a modified version of the UEQ was used. The questionnaire was adapted to focus on the participant's experience workflow/process, and questions focusing e.g. on the attractiveness or usability of the interface were omitted wherever possible. As such, we attempted to avoid the bias concerning the user interfaces which might be unfamiliar in some case, or not relevant in other cases where evaluating the efficiency of the process itself is the primary goal. The remaining 13 adjective pairs used in the questions were the same for all three use cases, though the phrasing of the question varied by use case (e.g. "Editing with metadata was practical/impractical" vs "Searching with metadata was practical/impractical").

The UEQ survey was complemented by further sets of Likert-type questions, relating to specific aspects of the platform, its usability and user experience, as well as brief open questions regarding the experience. These additional questions were about the quality of auto-generated metadata such as machine translation, speech recognition, face recognition and named entity recognition, as well as the quality of the subtitle spotting and segmentation and the effort involved in correcting them in the subtitling user stories.

[5] https://www.ueq-online.org/

5.3 Semi-structured interviews

In all but one case (i.e., *Interlingual subtitling – consumer reception* from Epic 6.11), a brief semi-structured interview was also carried out with each participant after completing the tasks to collect more detailed feedback on their experience, issues affecting the process and usability, and possible suggestions for future development and improvements. The interviews were recorded, then transcribed and anonymized. Thematic analysis [3] was then carried out on the transcripts. The responses by the participants were analyzed for positive vs. negative comments and specific issues raised by the participants, such as features impacting quality and usability or suggestions for improvement.

5.4 Process data collection

For the evaluation of user stories involving post-editing of intralingual and interlingual subtitles (i.e., Epic 6.11), subtitling process data were also collected to obtain information on how the use of ASR or MT output to be corrected (post-edited) affected the productivity and work processes of the subtitlers. In the third round, process data were collected in the form of total task times. The participants in use cases for intralingual subtitling and interlingual subtitle post-editing were asked to record the time taken to post-edit each file and to record any interruptions of their work during the task, if such occurred. They were then asked to report the task time in the post-task questionnaire to the nearest minute. The task times were used to assess the average temporal post-editing effort needed. For intralingual subtitling, task times were also compared to subtitling from scratch. For interlingual subtitling post-editing, task times were compared between the post-editing of different MT outputs.

As a proxy to measure the amount of post-editing done, edit distances were calculated comparing the original outputs and the post-edited subtitles. For intralingual subtitling the metrics used were Word Error Rate (WER) and Letter Error Rate (LER). For interlingual subtitle post-editing, the metrics were the word-level Human-Targeted Edit Rate (TER) and character-level TER (cTER), which were further used to compare edit distances between different MT outputs. Collection and analysis of process data is described in more detail in Sections 8.1.1 and 8.2.1. For further discussion of MT-based post-editing, see also deliverable D4.2 from WP4.

5.5 Focus group discussions and survey

The viewer studies for evaluating *Interlingual subtitling – consumer reception* from Epic 6.11 employed two data collection methods: focus group discussions and a questionnaire. The focus group is a research method where a small number of research participants is encouraged to discuss the topic of research, led by a moderator who steers the conversation in the appropriate direction with questions and comments. The participants are allowed to discuss the topic with each other, which can produce richer data than interviews, where participants simply answer questions [4]. The use of focus groups is a useful method for exploratory research and for developing more detailed research questions. They can be a way of brainstorming, uncovering a variety of subjective views and perspectives, and putting surface-level data into realistic, personal contexts. The focus group was therefore employed to scope attitudes and pinpoint specific areas of further inquiry. Note too that focus groups were also employed in the evaluation of the content description prototype (Epic 6.10).

Due to the necessarily small group sizes, focus groups do not provide generalizable quantitative data. Therefore, it is often useful to complement focus group data with a more quantitative approach, such as the questionnaire in this study. The questionnaire was used to gain a larger number of answers from respondents with more varied backgrounds. The questions were designed on the basis of early experiences from the focus groups. The UEQ survey used in other evaluations was not considered suitable for the viewer study, because the questions would have been too detailed for a viewer recalling a normal viewing experience. Furthermore, the purpose of the viewer study was to explore the comprehensibility and acceptability of MT subtitles, and the UEQ survey does not contain questions that address these topics. Instead, we designed a survey with a combination of multiple choice, 5-point Likert scale and open questions to gauge viewers' comprehension of the video clips, cognitive load caused by the subtitles, and their attitudes towards the clips and towards MT subtitles in general. The questionnaire and the focus groups can be seen as a way to triangulate each others' findings and to produce a more well-rounded view of the reception experience.

5.6 Materials used in the evaluation

In general, the audiovisual material used in the prototype evaluations discussed in this report was sourced from the data sets provided by consortium partners YLE and INA (as described in D1.2). These materials were ingested into the platform (discussed in D6.5, Section 6.1.1) and made available for a variety of purposes, including searching, export to video editing tools and for the application of subtitles.

In most cases, a further subset of this data was selected with the theme of European Parliament Elections. This theme covers a broad range of topics, appearing people and program types to make the evaluation tasks credible when compared to actual production use, while keeping the amount of content small enough to be manageable in terms of running multiple analyses on the evaluation materials multiple times during development. The elections theme was interpreted loosely to catch a good variety of programs covering Europe, politics, economy, culture and not to focus too tightly on e.g. election debates and election results reporting only.

The materials used for each evaluation track are described in detail in each of the following sections.

6 Evaluation of Epic 6.5: Editing assistance using multi-modal and multi-lingual metadata

In this evaluation round, process data was collected from (news) video editors working with multimodal automated metadata (ASR, face recognition, NER and machine-translated metadata) (user story 2.1.5) and machine translations (user story 2.1.6). This evaluation was the second round organized for these user stories and a follow-up of evaluations done in December 2019.

The purpose of collecting and analyzing user data was to determine a) how automatically generated metadata and b) how automatic translation of transcripts from raw footage affect the work of video editors. In this evaluation round, the test-setup included an improved set of metadata and a slightly different group of participants. The purpose was to collect more information on how auto-generated metadata could be utilized in a professional video editing process.

The evaluation was conducted in seven individual test sessions that were carried out between June 23rd and July 6th 2020 at the YLE premises.

Video editors' subjective evaluations of the usability of automatically extracted metadata, ASR and MT for these purposes were collected using the UEQ survey and semi-structured interviews (cf. subsections 5.2 and 5.3).

6.1 Motivation

Through this evaluation, we wish to learn to what extent the editing process is made more efficient due to the auto-generated metadata that is made available directly in the non-linear editing software used for craft video editing. We felt a second evaluation was needed in this case, as during the previous evaluation round, the users somewhat questioned the added value of enriched metadata for their editing workflows. Some of the participants were not fully convinced of its usability, whereas some felt that content retrieval and organization activities based on this metadata (i.e., user story 2.3.1) were not primarily part of their core work (cf. D6.6, p24). In this second evaluation round, an improved set of metadata was used as well, and the participants' group was changed. The motivation was to determine whether the assessments on the auto-generated metadata's usability would change with these test setup adjustments.

6.2 User test setup

As part of the description of the user test setup we provide insights into which audiovisual material was used, who participated in the evaluation, how experiment data collection was done, and finally, which tasks users were asked to execute as part of the evaluation.
For the most parts, the user test setup was kept the same between this second and first evaluation for this Epic 6.5. More details on this setup were discussed in D6.6. We highlight the differences in this second evaluation round.

6.2.1 Material used and testing environment

For this evaluation, the same single video was used as in the first evaluation, namely a 2hr recording from the MeMAD catalogue (cf. D6.5 and D1.2), representing:

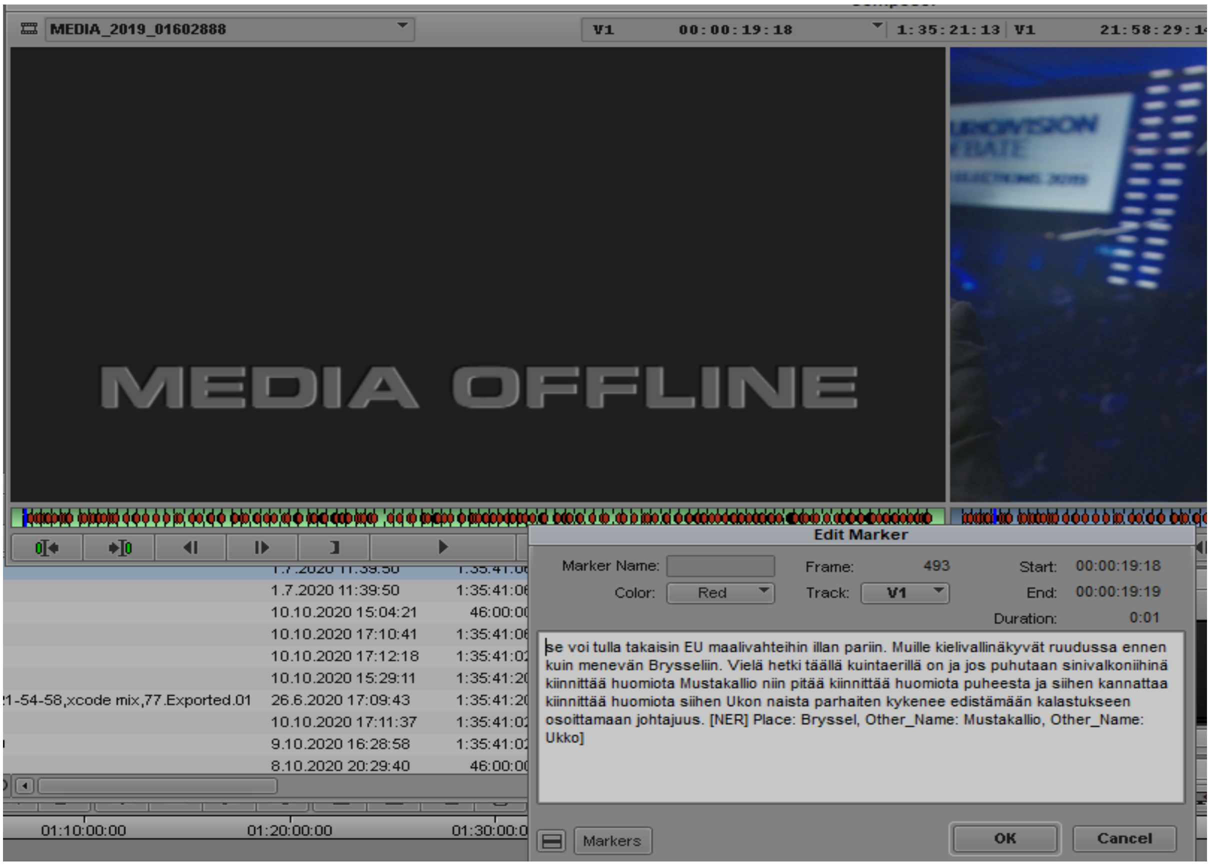
1. various topics;
2. featuring various but a limited number of people;
3. featuring languages suitable for testing MT.

The selected program is one of the lead candidate debates from 2019 European Parliament elections[6]. The 2-hour debate contains thematic sections on economics, immigration, taxation, etc., each ca. 10 minutes in length. The main language of the debate is English, with an introduction in Finnish and some sections in French, which were used to test the execution of the MT user story. Due to the changes in the evaluation user panel, this content was still unfamiliar to participants and it was used as raw material which made it possible to create a simulation of a genuine video editing process.

We reused the same source metadata generated with this clip as before, but we did improve the way the metadata was imported into the editing environment (as also discussed in Section 6.4 of D6.8). The improvement of the enriched metadata was mainly to improve searchability for face recognition, audio transcripts, and machine translation metadata. The aim was also that the results coming from face recognition and transcripts could be merged in one metadata marker instead of having to retrieve them separately (cf. user story 2.1.5). Some re-adjustments were also made since the previous evaluation as it was also noted in the evaluation plan that the length of the metadata markers should be shortened since they had been quite long last time and rather difficult to read.

During the previous evaluation, there had been some complaints regarding the test setup. The sessions then were carried out using a journalists' edit-workstation housed in an office environment. It lacked some key essentials such as a color-coded Avid software keyboard and an additional video monitor. Nor did the location have any acoustic features corresponding to an authentic production environment. This second evaluation round hence moved to a real news production environment that resolved these complaints. Figure 5 depicts this new representative working environment used for this evaluation round.

[6] Also available online at: : https://areena.yle.fi/1-50141056



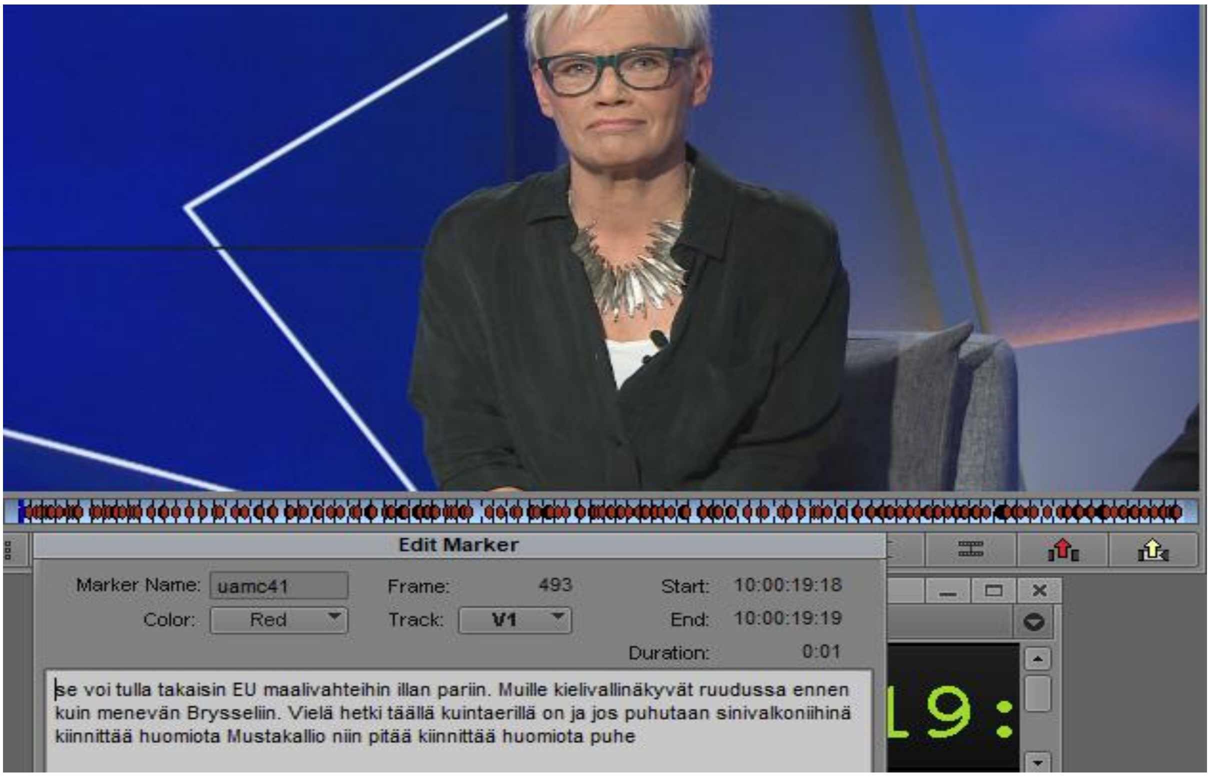
1.7.2020 11:39:50 1:35:41:06
10.10.2020 15:04:21 46:00:00
10.10.2020 17:10:41 1:35:41:06
10.10.2020 17:12:18 1:35:41:02
10.10.2020 15:29:11 1:35:41:20
1-54-58,xcode mix,77.Exported.01 26.6.2020 17:09:43 1:35:41:20
10.10.2020 17:11:37 1:35:41:02
9.10.2020 16:28:58 1:35:41:02
8.10.2020 20:29:40 46:00:00

Edit Marker

Marker Name: Frame: 493 Start: 00:00:19:18
Color: Red Track: V1 End: 00:00:19:19
Duration: 0:01

se voi tulla takaisin EU maalivahteihin illan pariin. Muille kielivallinäkyvät ruudussa ennen kuin menevän Brysseliin. Vielä hetki täällä kuintaerillä on ja jos puhutaan sinivalkoniihinä kiinnittää huomiota Mustakallio niin pitää kiinnittää huomiota puheesta ja siihen kannattaa kiinnittää huomiota siihen Ukon naista parhaiten kykenee edistämään kalastukseen osoittamaan johtajuus. [NER] Place: Bryssel, Other_Name: Mustakallio, Other_Name: Ukko]

Figure 3: View of updated AME metadata provided to users in the Avid Media Composer editing software in the 2020 editing assistance evaluations (1/2).



Edit Marker

Marker Name: uamc41 Frame: 493 Start: 10:00:19:18
Color: Red Track: V1 End: 10:00:19:19
Duration: 0:01

se voi tulla takaisin EU maalivahteihin illan pariin. Muille kielivallinäkyvät ruudussa ennen kuin menevän Brysseliin. Vielä hetki täällä kuintaerillä on ja jos puhutaan sinivalkoniihinä kiinnittää huomiota Mustakallio niin pitää kiinnittää huomiota puhe

Figure 4: View of updated AME metadata provided to users in the Avid Media Composer editing software in the 2020 editing assistance evaluations (2/2).

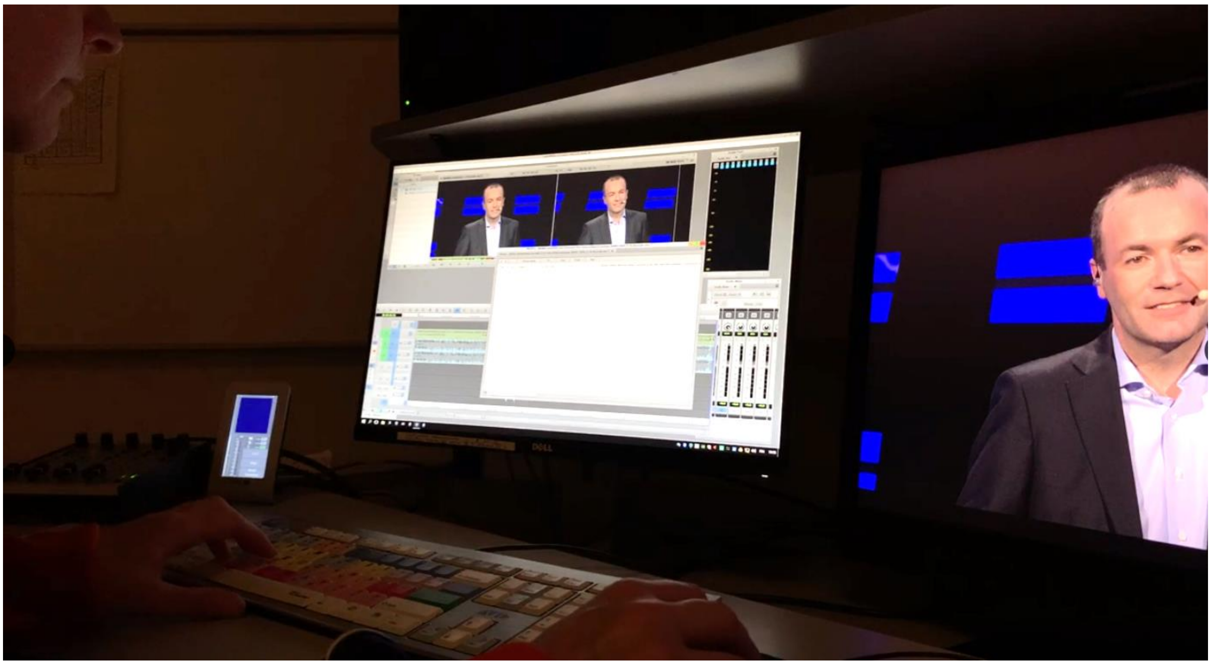


Figure 5: Depiction of the updated video editing evaluation environment used for Epic 6.5.

6.2.2 Participants

In the previous evaluation round, process data was collected from a group of video editors working mainly on YLE's weekly programs. This time, the data is collected from a new target group, a group of four video editors and two cameraman/editors all working in YLE's news department. The question of whether the original choice of the participant group should be reassessed rose after the previous evaluation round as the data collected from the test sessions showed that pure video editors do not always see retrieving different fragments from extensive video material unknown to them as being part of their core work (cf. D6.6, p 24). Such reactions were uttered due to the fact that the work of a video editor who works mainly on weekly programs rarely involves any retrieving of material while that kind of task is primarily seen as being part of a director's or journalist's responsibilities.
To help mitigate this mismatch, news video editors were chosen to be the new target group, as the building process of a news story often requires retrieving content from several program sources. In addition, news video editors usually work in a fast-paced production environment where they are expected to make independent decisions, which was thought to affect the way they would approach the process of material retrieval, handling and gaining insights into the material's contents.

As with the previous evaluation round, the Avid Media Composer professional craft video editing software was used as the evaluation tool. Again, all participants were familiar with the editing software. All participants took part in the evaluation of both user stories (i.e., 2.1.5 and 2.1.6).

None of the participants were fluent in French, so they relied on the machine translations available for the French part of the tasks. As a new feature in the test setup, the participants were allowed to use Finnish as their search language. This meant that search results could also be viewed in Finnish (the only exception being that the metadata tags created from named entity recognition were in presented in English).

6.2.3 User data collection and tasks

The experiments for editing assistance data collection were arranged at YLE premises in July 2020. As mentioned above, the editing assistance tasks were carried out using Avid Media Composer. The news editors had access to the internet and most other resources normally used in their work.

A previous pre-task questionnaire was re-used to collect background information from the participants (cf. Appendix A in D6.6), and a post-task questionnaire was used to collect subjective assessments of the editing experience and the quality of the available metadata. After the completion of the tasks, a brief semi-structured interview was also carried out to collect more detailed feedback regarding problems in the workflow and the participants' views on potential improvements.

Task summary

The evaluation sessions were designed to reflect a typical video-editing process in YLE's News department and when forming the editing tasks, the aim was to create a setup that resembled a typical editing session that normally involves material browsing i.e. searching for pre-defined specific topics and/or people in the freshly arrived video content.

In the editing task, the news video editors were asked to make a short video collage that includes all the predetermined scenes listed on the task list. After the assignment was given, the editors were reminded that the outcome would not be judged by any aesthetic criteria, but that the purpose was only to gather information on how well the relevant elements mentioned in the task list could be found using the enriched metadata.
No explicit time limit was given for the tasks, rather, the participants were instructed to work at their own pace. In all, each evaluation sessions lasted about two hours at a time. Before each session, the participants were introduced to the MeMAD project and explained in general terms what technologies had been used to produce enriched metadata. After the brief introduction, the participants began to complete the tasks independently in the order listed. The list of tasks given to participants was the same as in the previous evaluation and can be found in D6.6 under Appendix A.

6.3 Analysis of user data

6.3.1 User Experience Questionnaire

In this evaluation, as in the previous round, the UEQ questionnaire was again used after the editing task to collect the participants' subjective evaluations of the experience of using metadata. Figure 6 shows the averages of the six participants' responses to each of the questions on a scale from -3 to +3. Averages between -0.8 and +0.8 (marked with dashed lines in the figure) are considered neutral, with values exceeding this limit deemed negative or positive.

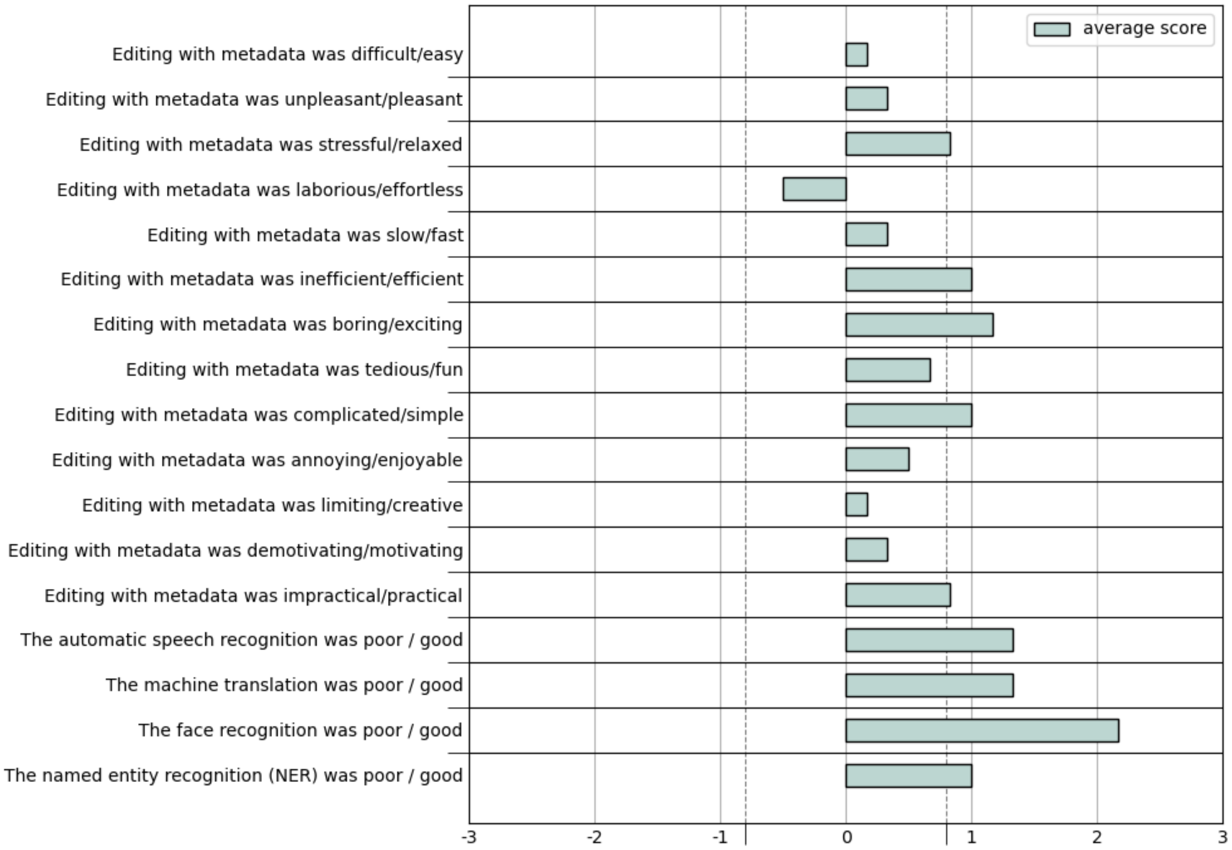


Figure 6: Average UEQ scores for Epic 6.5: Editing assistance, 2020 evaluation.

On average, the responses are neutral or somewhat positive. The most positive values are seen for the characterizations “exciting”, “simple”, “efficient”, “relaxed”, and “practical” of the presented editing process. The other adjective pairs remain in the neutral range, tending toward mildly positive. Only the evaluation for the adjective pair “laborious/effortless” tends toward negative, although it also remains within the neutral range.
In addition to the adjective pairs describing the editing experience, the participants were asked to assess the quality of different types of metadata on a similar scale (poor/good). All the metadata types are evaluated positively. The highest average is seen for face recognition, followed by speech recognition, machine translation and finally named entity recognition. Compared to last year’s UEQ scores, we see a slight improvement, especially with regard to the scores of the perceived quality of individual metadata modalities provided.

Summary of comments in the user experience questionnaire

In the post-task questionnaire, the participants were also asked to give short written comments on their experience of the editing process and the quality of the metadata. Two of the six participants (01, 04) commented that finding the correct place in the long video was much faster with the help of the metadata than otherwise. They noted that without the metadata (and without specific timecodes provided by the journalist) one would need to watch the video and rely on memory, fast forwarding and rewinding, or hope that they could find a description of the video content online. On the other hand, three participants (03, 05, 06) stated that finding the correct places was difficult. Reasons mentioned for this difficulty appeared to focus mainly on the features of the Avid software and the way it displays metadata and methods available for searching through metadata markers, which the participants did not consider useful. Some of the metadata also appeared not to have transferred correctly in the import. One participant (02) merely noted that the task was somewhat challenging but interesting, and that the metadata would need to be more accurate. Overall though, the participants characterized the metadata quality as reasonably good or sufficient. Of the specific metadata types, two participants (01, 04) mentioned that face recognition worked well and was helpful because it removed the need for the editors to search for images themselves. Three participants (01, 03, 04) commented on the machine translation, saying that it appeared comprehensible and sufficient for this purpose. Speech recognition and named entity recognition were not explicitly mentioned in the written comments.

6.3.2 Feedback from interviews

For this evaluation, the following questions and structure were used for the semi-structured interviews:

1. How did you feel about the editing tasks and why is that?
2. Was there anything positive/negative about the tasks? (Based on first answer; if their feelings are negative, ask about anything positive.)
3. What features of the provided data and the content impacted the editing the most, in good and bad?
4. Did you notice any differences in the data or transcripts?
5. What is your editing process usually like with content like this?
6. How did the use of MT/ASR impact your own work process? And the other metadata?
7. Could you imagine using MT/ASR as a tool?
8. How should it be improved?

In order to understand the user experience (the participants using MeMAD metadata functionalities in a video editing task), the interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements.

Like in the previous evaluation round described in D6.6 and in order to understand the user experience, the interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements. Interview questions 2 and 3 turned out to overlap. The number of interviewees being limited (N=6), we did not conduct any quantitative analysis of the interviews.

Most participants (4 from 6) found the experience pleasant and interesting; however, due to the tool, the central task was searching, which was assessed only by two participants in positive terms. According to them, it was quick and easy and the search result was

adequate, not too big or detailed, and it was easy to find the content even in a long material. One of the two could not mention any negative aspects.
The biggest problem seemed to be the Avid editing software. Three participants explicitly claimed that it was not integrated with the metadata's functionalities and caused technical problems. They regretted that searching was difficult, clumsy and slow. However, one of the critical participants saw the functionalities in Limecraft Flow's project library (used as the basis for the MeMAD integrated platform) and described the metadata and search features as "amazingly fine". One participant criticized the display of the metadata in one single column, and two suggested using different colors in the elements or markers. On the negative side, also search terms were mentioned. Three participants regretted that only one word could be used. Accordingly, the possibility of using several search terms and of starting for example from a certain time code was suggested. Further suggestions for improvement included the development of search features and better integration with Avid.

Overall, the participants found it difficult to assess the different functionalities and to know which functionality led to the right result (questions 4, 6 and 7). However, they acknowledged the usefulness of face recognition, ASR and MT in searching and finding material, and half of the participants regarded the quality of face recognition and MT as good (other participants did not assess the quality). Especially MT made the search easier and faster, but also allowed independent working. At least a translation into English would be necessary. One participant mentioned the risk of relying on the automatically generated data only which gives the machine more control over what will be chosen. Another risk mentioned was to choose the opposite opinion instead of the right comment: when editing interviews, for instance, knowing the topic is not enough but the editor has to know what speakers say in order to choose appropriate content.

One participant considered NER as handy, while others found it difficult to assess it or did not comment on it. The answer to question 4 ("Did you notice any differences in the data or transcripts?") would have required a careful reading of the transcripts, instead, the participants concentrated on single words. As such, not conclusive answers were obtained for this question.

All participants could imagine using the metadata functionalities tested in their work, especially if the metadata had some improvements. However, all of them emphasized that in the normal workflow, searching is not done by them but by the journalists from whom they get the time codes. Only in the rare cases in which journalists are busy or if they are working on live broadcast, the editing assistant can help in searching. Two participants saw a risk in the new tool: it could shift the journalist's work and at the same time journalistic decision-making power to editing assistants and increase their workload.

Similar critique towards the feasibility of the search task for editing professionals was raised in the previous evaluation round (D6.6). Nonetheless, we decided to conduct this last evaluation because of a modified context and tasks which are better suited to the editing profession (i.e., with the modified group of participants). One editor participated in both evaluation rounds (nr. 3). The context in which we tested the developed functionalities was news editing: this environment is more apt to this kind of editing task which requires fast processing and production of content. The data were the same (EU debates) but the updated tasks, news editing, were now more realistic. The tasks were carried out in an authentic environment at the news section of YLE and there was no time limit, compared to the “unnatural” office test environment in the last round. The news content is typically multilingual, and in this final evaluation round we provided the editors automatic translations into Finnish, whereas last round provided translations only into English. Furthermore, we included a new professional type, the news video editor, to gather information about how the search task fits into their work processes.

6.4 Discussion of the user feedback and data gathered

Both the data gathered from the evaluations and the informal discussions with the participants seem to confirm that the video editors, while they do perceive a use for the AME metadata, do not see if adds value to their work, or at least they are not very convinced. Rather, they tend to see all content retrieval operations where the metadata do serve their purpose as belonging to the journalist, who is responsible for gathering all necessary content before the actual edit process. One possible reason for this could be that the idea of video editing being a technical process seems to be firmly established in the industry and is not easy to change whereas journalist’s work is seen as being more information-based. This has probably affected the idea of which tasks belong to the video editor and which do not.

On the other hand, video editors work is known for being very fast-paced and tool-centric. In YLE's news production, they mainly edit multiple news stories in a single shift and besides editing, the making process typically also involves sound mixing. The work is by nature a mixture of technical functions and creative solutions, and to succeed in the job requires reliable tools that can be controlled intuitively. Thus, it is likely that the slightly reluctant attitude of video editors towards the use of auto-generated metadata is largely due to the fact that at present its use seems to be poorly integrated into their own workflows. As soon as the enriched metadata can be used in an interface seamlessly integrated with the editing software, which allows the metadata to be used dynamically and in no way slows down the fast-paced editing process, users will most likely be more positive about it.

6.4.1 Technical challenges

The way Avid Media Composer handles imported metadata is somewhat limited. This is especially evident when the imported data contains long transcripts generated by ASR and MT.
In Avid, the transcribed text is not synced with the original audio and therefore it can't be used as dynamically as in Limecraft Flow, instead all imported text must be included in chapters, the length of which must be specified in advance.
After importing chapters into Avid, they need to be opened in a Marker tool that presents each metadata record in a single line view. This was considered rather limiting since the evaluation concerned multimodal metadata generated in different sources that consisted of tags and long pieces of transcribed and/or translated texts.

During the first evaluation the participant criticized that in order to view the markers imported into Avid, they appeared by default in the sequence view which is not ideal for the normal workflow of a video editor, where all the ingesting should be done in the playback window. The participant tried to work around the limitation by first copying the source material in Avid's Media Bin and then copying the metadata from the original media to the copied clip. This way the imported chapter could also be seen in Avid's playback window. However, it appeared later that operating with the imported chapter markers other than in sequence mode, part of the metadata was not recognized.

6.4.2 Alternative video editing systems

It is possible that other professional video editing systems available, such as Adobe's Premiere Pro or Black Magic's Davinci Resolve, could have handled the imported metadata more flexibly since it's known that those companies follow a more open policy concerning third-party software integrations. However, since YLE uses Avid Media Composer as a corporate level video editing system, no other video editing software was tested in the evaluation. At the same time, many alternative tools are built around similar non-linear editing concepts with limited support for temporal metadata besides those used purely to support the functional cutting and editing process.

6.4.3 Possible future applications

During some open conversations with the participants, it was noted that concerning possible future applications, it might be beneficial if the automatically provided metadata could be attached with live recordings, for example, with international news feeds that are continuous program streams in which news topics are run sequentially. Broadcast companies such as YLE are recording those feeds provided by international news agencies around the clock or receive fresh content from camera crews working on the field, and sometimes finding a specific scene or object in this material can be very time-consuming. The sentiment was that in fast-paced news production, such content enrichment could play a more central role than for the existing archive material.

7 Evaluation of Epic 6.3: Searching and browsing for ingested and archived content

In section, we discuss a second evaluation round in which process data was collected from media production and archive professionals searching for information and media from media archives. The purpose of collecting analyzing process data was to determine how automatically generated and semantically linked metadata and machine translations affect video searching on a) program/item level (user story 2.2.1) and b) sub-program level such as segments (user story 2.2.2). This evaluation was the second round organized for these user stories and a follow-up of evaluations done in January of 2020.

The evaluation sessions took place in November and December 2020 with a group of media professionals working for YLE and KAVI (the Finnish National Audiovisual Institute). In this revised study, professional media archive users performed a series of media and information retrieval tasks using different combinations of automated metadata, incl. ASR and MT outputs, recognized named entities, face recognitions and scene and object classifications.

As with the previous iteration, this evaluation was designed to reflect everyday use of professional media archives. Typical search scenarios aim to find either a) media content to be re-used or b) information which is based on or derived from the media collection and its database. Re-use of media content is typically either re-publishing existing content such as programs or extracting parts of programs (clips) to be used as raw material in new productions. Information derived from the media collection is for example listings of programs by a certain person, which can be used in background research for journalism.

This evaluation focuses on re-use of content, which is closer to the project scope as it focuses typically on the content itself instead of administrative metadata such as lists of people who have contributed to programs.

For the media searching tasks, the typical outcome is a selection of potential clips or programs, out of which video editors can choose what they finally end up using in their productions. Some tasks have of course a single correct answer, such as "Find me the clip where Kennedy says 'Ich bin ein Berliner'", but for most cases the expected result is not unambiguous. Rather a short list of "good enough" candidate clips or programs is the search result which can then be refined iteratively by media archivists and their clients such as video editors or journalists. For this evaluation, the initial search result was the expected outcome and the iterative refinement of task criteria is seen as out of scope. We did revise the provided search tasks to better align with realistic content retrieval scenarios encountered in real working scenarios. These revisions were based on feedback from the previous evaluation, and the availability of more AME metadata, especially concerning the visual domain, which makes more extensive searches possible.

As before, the test panel's subjective evaluations of the usability of auto-generated metadata were collected using the UEQ survey and semi-structured interviews. Given restrictions in the test setup due to the COVID-19 pandemic (cf. also Section 5.1), no more think-aloud verbalizations were collected from the participants.

7.1 Motivation

The aim of this evaluation was again to learn whether users can succeed in retrieving content items, and relevant segments within these content items, from an archive when they have been enriched by various metadata auto-generated by MeMAD content analysis components, including ASR, named entity recognition, face recognition, etc. Using a set of content retrieval tasks, we gauge how successful the test panel is at finding relevant items given a set of search criteria. Additionally, we want to learn to what extent the auto-generated metadata can substitute human-curated 'legacy' metadata in this content retrieval process. To take full advantage of the metadata produced, searching is done using the Limecraft Flow search interface which provides the front-end of the MeMAD prototype platform. For this final evaluation round, we applied the latest metadata configurations regarding machine translation, face recognition and visual scene classifications to the evaluation test set. An important question was hence to gauge whether the updated set of auto-generated metadata work more effectively than on the previous test rounds. The set of search tasks was reformed to allow more search methods to be used and to provide more diversity into the final evaluations.

7.2 User test setup

As part of the description of the user test setup we provide insights into which audiovisual material was used, who participated in the evaluation, how the experiment data was collected, and finally, which tasks users were asked to execute as part of the evaluation. Many of these aspects remained unchanged from the previous evaluation, as discussed in Section 7 of D6.5. Here, we highlight the evolutions and differences between the two evaluation studies.

7.2.1 Material used

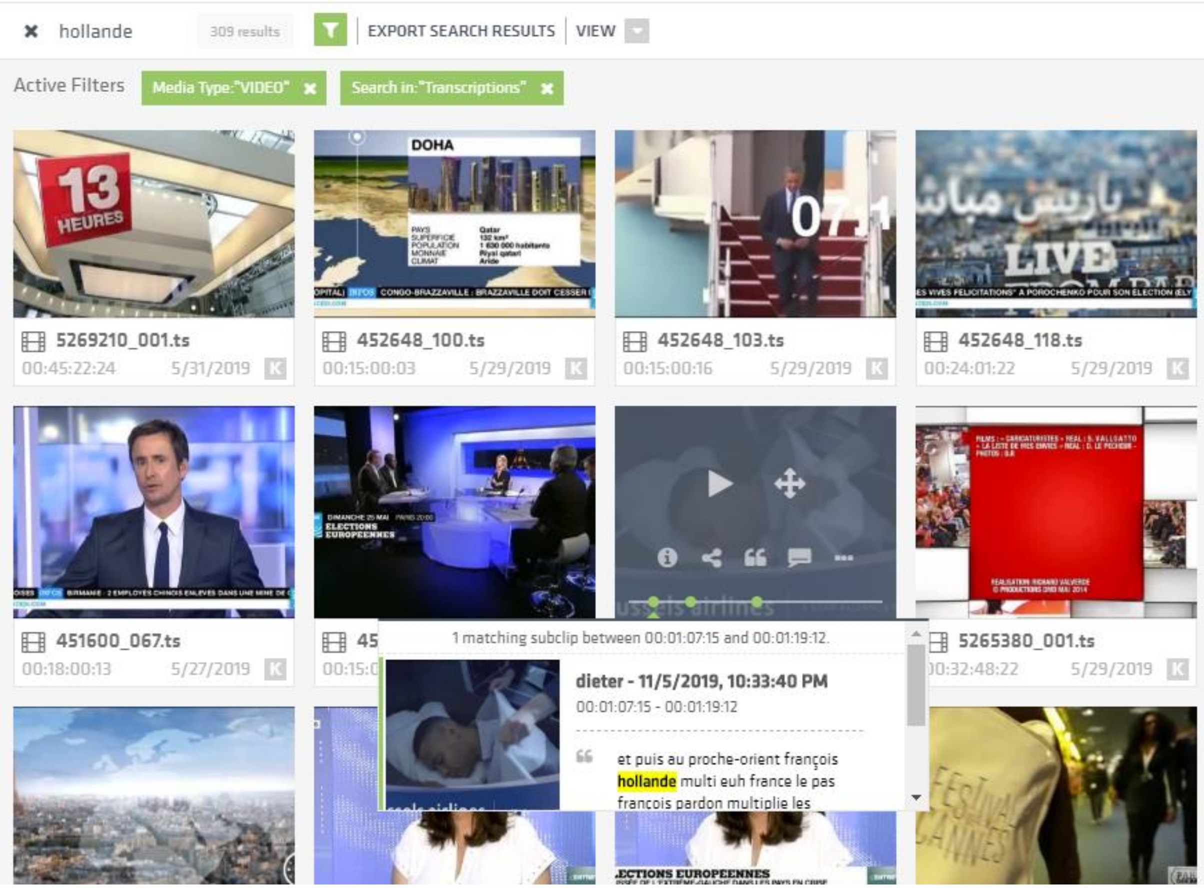
The same collection of 408 video clips (of approximately 210 hours of media content) was used as in the previous evaluation round (cf. D6.6, subsection 7.2.1). As before, this subset of the entire MeMAD catalogue of content was made available through a separate sub-collection in the Limecraft Flow prototype user interface such that users were constrained in the content they could retrieve.

With respect to metadata enrichments, many of the original metadata inputs used in the first evaluation were retained (cf. to D6.6 for more details) in this second study, but many metadata were added, which we list here:

- All but a few items were previously audio-transcribed by MeMAD services. Items that were previously transcribed using Lingsoft's Finnish or Swedish ASR were redone to take advantage of better speaker diarization and subsequent transcript splitting according to speaker turns (cf. D6.8, section 6.2.1, and D2.3). This delivered smaller chunks of search results that allowed users to better isolate relevant segments. As before, no post-editing was performed on the transcripts. An illustration of ASR transcript search results is depicted in Figure 7.
- All items with speech transcripts in French or Finnish had been previously machine-translated into English as the common language. Due to the availability of additional language translation pairs in the MT services available from

University of Helsinki (cf. D4.3), this was extended with additional translations from French and English into Finnish for this evaluation round. Especially for INA-provided content, this meant a much larger coverage of machine translated speech in the native language of most of the evaluation participants.
The availability of these MT transcripts also allowed for an indirect evaluation of the cross-lingual search mechanism we described in Section 4.1 of D4.4.

- All items were indexed with the custom-trained instance of the EURECOM face recognition service (cf. D6.8), which was trained to recognize a selection of 48 persons that commonly occurred in the test set, including prominent EU elections candidates and states leaders in Europe. Figure 8 illustrates an example.
- Visual text detection using optical character recognition (OCR) was also applied to the test set collection (cf. D6.8, Section 6.2.1). Each element of detected text was added as a logical subclip with, as its description, the text found in the image along with the spatial coordinates where the text is located in the video image. As such, the text was indexed and could be found by users looking for these terms directly, as shown in Figure 9 with OCR results for the term “finnair”.
- Lastly, visual scene classification metadata was added thanks to the integration with the AALTO PicSOM framework (also discussed in D6.8, Section 6.2.1). On a per-shot level, each video segment was classified with one or more of the 397 classes from the SUN397 Scene Categorization Benchmark database [5]. An example search result for the term “cows” is depicted in Figure 10.



hollande
309 results

Active Filters: Media Type:"VIDEO" Search in:"Transcriptions"

5269210_001.ts
00:45:22:24 5/31/2019

452648_100.ts
00:15:00:03 5/29/2019

452648_103.ts
00:15:00:16 5/29/2019

452648_118.ts
00:24:01:22 5/29/2019

451600_067.ts
00:18:00:13 5/27/2019

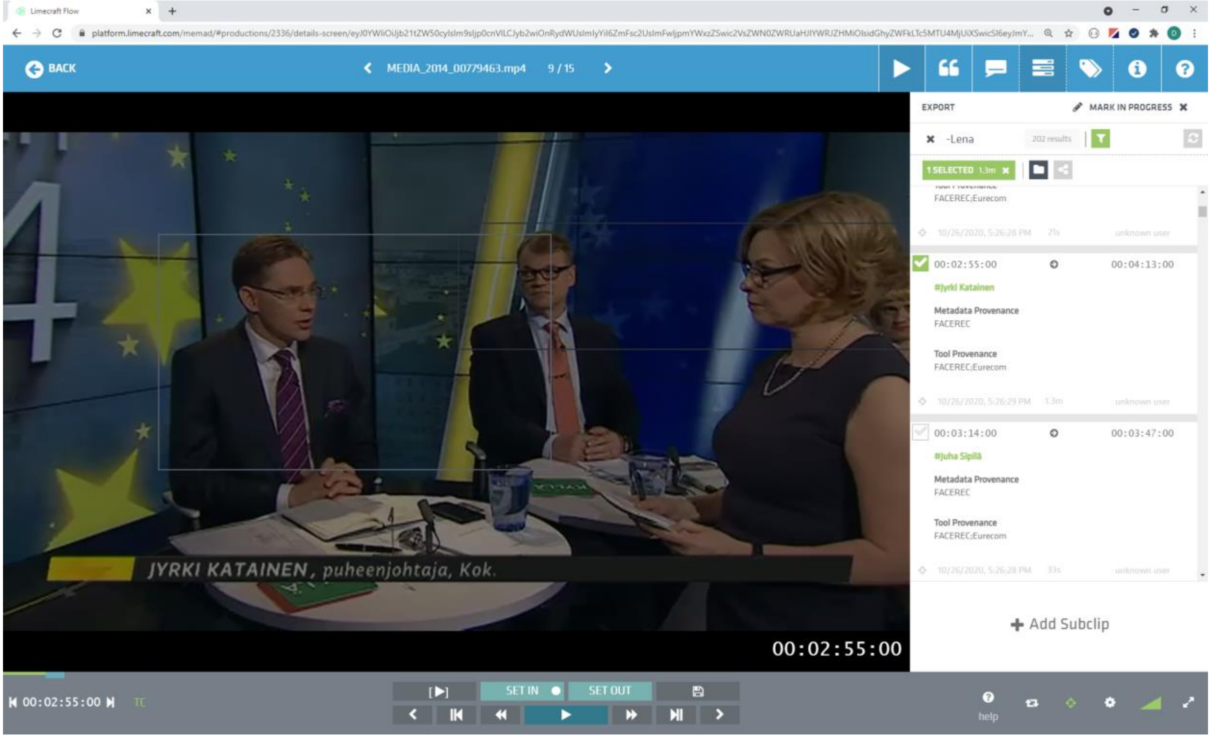
1 matching subclip between 00:01:07:15 and 00:01:19:12.

dieter - 11/5/2019, 10:33:40 PM
00:01:07:15 - 00:01:19:12

et puis au proche-orient françois hollande multi euh france le pas francois pardon multiplie les

5265380_001.ts
0:32:48:22 5/29/2019

Figure 7: The platform library shows search results for the term "hollande", with clip and audio transcript part matches in the search results.



-Lena 202 results

1 SELECTED 1.3m

FACEREC;Eurecom

10/26/2020, 5:26:28 PM 21s unknown user

☑ 00:02:55:00 00:04:13:00

#Jyrki Katainen

Metadata Provenance
FACEREC

Tool Provenance
FACEREC;Eurecom

10/26/2020, 5:26:29 PM 1.3m unknown user

□ 00:03:14:00 00:03:47:00

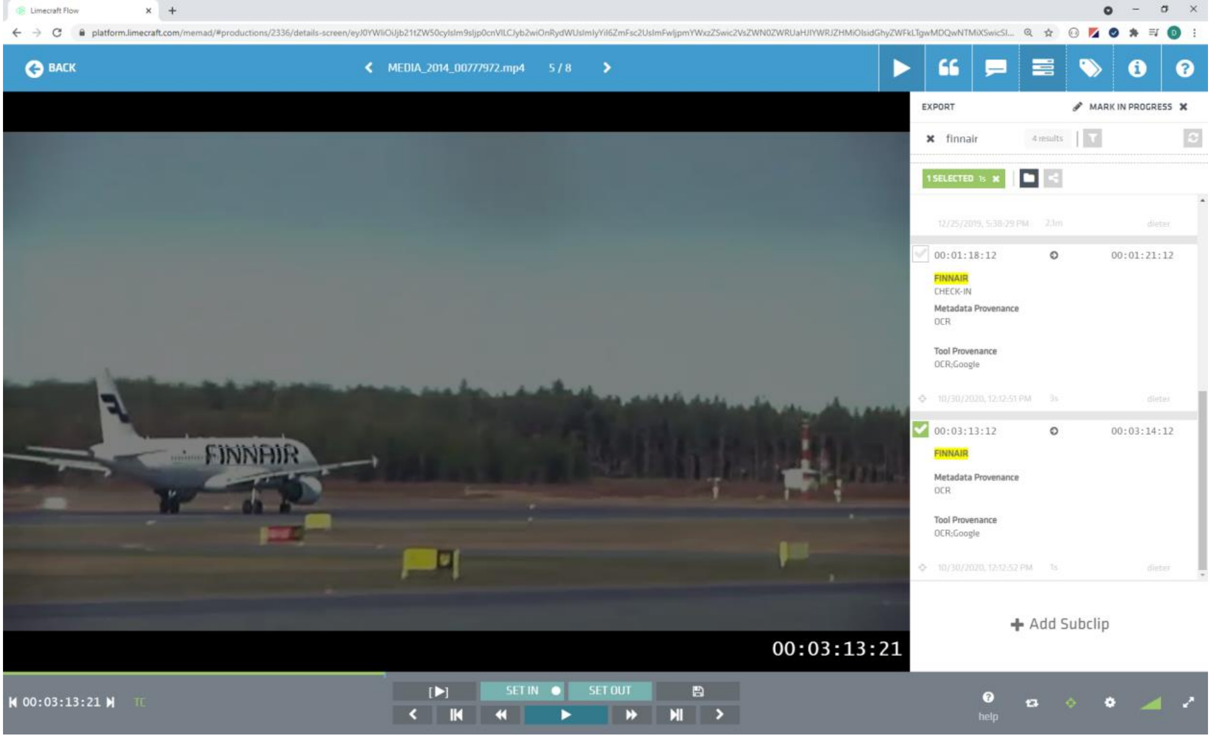
#Juha Sipilä

Metadata Provenance
FACEREC

Tool Provenance
FACEREC;Eurecom

10/26/2020, 5:26:28 PM 33s unknown user

Figure 8: Face recognition results, incl. a match for "Jyrki Katainen", as confirmed by the on-screen text.



finnair 4 results

1 SELECTED 1s

12/25/2019, 5:38:29 PM 2.1m dieter

□ 00:01:18:12 00:01:21:12

FINNAIR
CHECK-IN

Metadata Provenance
OCR

Tool Provenance
OCR;Google

10/30/2020, 12:12:51 PM 3s dieter

☑ 00:03:13:12 00:03:14:12

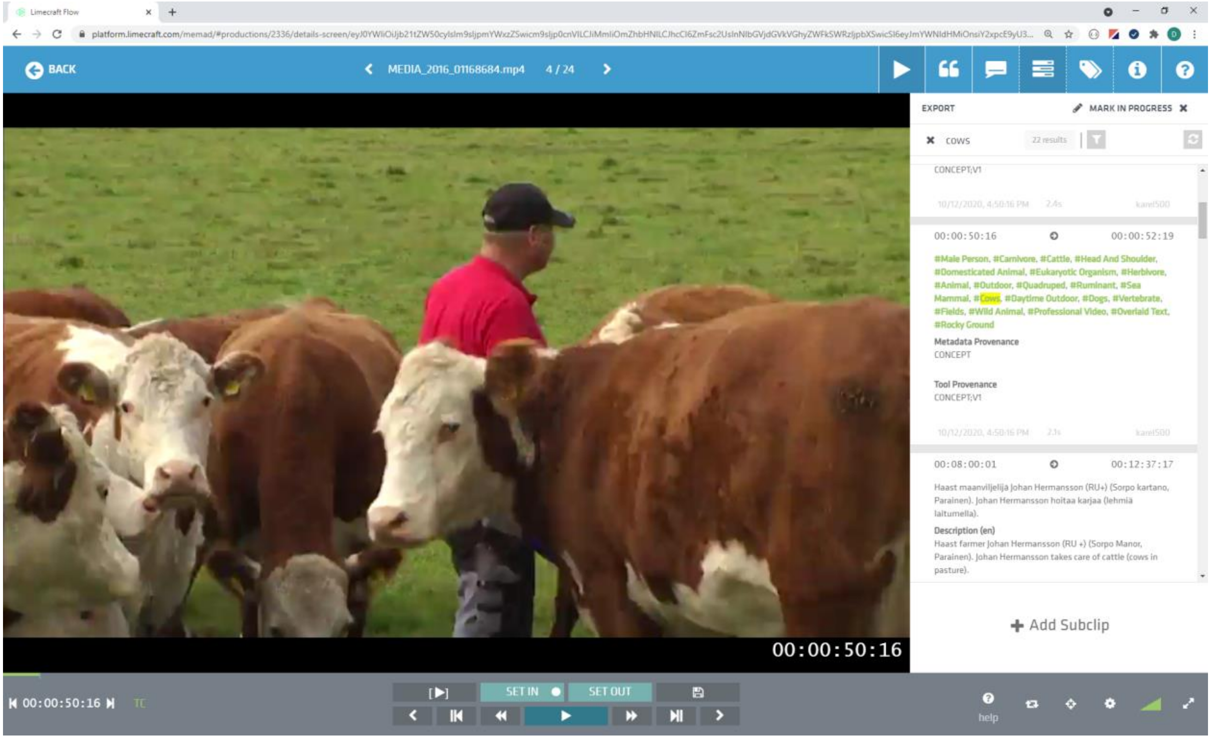
FINNAIR

Metadata Provenance
OCR

Tool Provenance
OCR;Google

10/30/2020, 12:12:52 PM 1s dieter

Figure 9: "Finnair"painted on the side of this airplane was detected by the OCR process, and added to this video item as a descriptive metadata element.



CONCEPT;V1

10/12/2020, 4:50:16 PM 2.4s karel500

00:00:50:16 00:00:52:19

#Male Person, #Carnivore, #Cattle, #Head And Shoulder, #Domesticated Animal, #Eukaryotic Organism, #Herbivore, #Animal, #Outdoor, #Quadruped, #Ruminant, #Sea Mammal, #Cows, #Daytime Outdoor, #Dogs, #Vertebrate, #Fields, #Wild Animal, #Professional Video, #Overlaid Text, #Rocky Ground

Metadata Provenance
CONCEPT

Tool Provenance
CONCEPT;V1

10/12/2020, 4:50:16 PM 2.1s karel500

00:08:00:01 00:12:37:17

Haast maanviljelijä Johan Hermansson (RU+) (Sorpo kartano, Parainen). Johan Hermansson hoitaa karjaa (lehmiä laitumella).

Description (en)
Haast farmer Johan Hermansson (RU +) (Sorpo Manor, Parainen). Johan Hermansson takes care of cattle (cows in pasture).

Figure 10: Shot-level classification results using the SUN397 database ontology.

For these evaluations, we re-used Appendix D of D6.6 as a succinct introductory manual and guidelines provided to the test panel to guide them through the tasks of this evaluation. In this guide we provided details about how they can define complex searches, filter down on results and make preset selections on which metadata is considered for generating search results.

To ensure that users can properly determine which search results are obtained using which metadata, we introduced a metadata provenance scheme such that each generation of metadata could be individually tracked and enabled or disabled, despite being sourced from a different origin. For example, metadata that exists in a different language – speech transcripts are the obvious example – cannot be easily distinguished purely on their appearance in the GUI (they both contain similar speech fragments). By adding provenance metadata fields, we can determine how the metadata was produced (e.g., from ASR, MT or Manual creation), and we were able to setup the GUI such that only those metadata results from a given provenance were retained.

7.2.2 Participants

The test panel consisted of seven participants. Five of them were archival professionals, i.e. archival journalists and catalogers working for YLE and KAVI. The two other participants had a close relationship to the archival process, one being a journalist in YLE's news department and the other working for KAVI as a technical specialist.
A previous pre-task questionnaire was re-used to collect background information from the participants (cf. Appendix B in D6.6).

Out of the five participants, none were fluent in French, so for the tasks dealing with materials in French they had to rely on the machine-translated transcripts for content retrieval.

7.2.3 User data collection and tasks

The evaluation sessions were carried out between November 13nd and December 3rd 2020. Due to the COVID-19 restrictions, each session was conducted in a one-to-one online meeting, with participants using Limecraft's search interface environment on their own computer and being asked to share their screen with the test coordinator. In this way, the participant's activities on the screen as well as the dialogue between the participant and the coordinator could be recorded as part of the research data. This was not always a smooth process, as the the computer hardware and network speed used by participants were not always optimal regarding the remote sessions as well as some of the participants were facing difficulties in sharing their desktop in Google Meets online meeting. This also led to an forsaking of the strict think-aloud process in this evaluation, as it was found that it could not be combined reliably with the user screen-sharing, remote instructions and actual content retrieval tasks.

As before, the test panel performed a series of media and information retrieval tasks using different combinations of automated metadata. Additionally, the participants had access to the internet and most other resources normally used in their work.

A pre-task questionnaire was used to collect background information from the participants, and a post-task questionnaire was used to collect subjective assessments of the content retrieval experience and the quality of the available metadata. After the completion of the tasks, a brief semi-structured interview was also carried out to collect more detailed feedback regarding problems encountered during the search process and the participants' views on potential improvements.

Each evaluation session lasted a total of 3 hours, including a 30-minute preparation period during which general information about the MeMAD project was provided, as well as a brief description of the key technologies that were used to produce metadata automatically.

The participants progressed on the task list by completing each search task twice first in a filtered metadata mode where every search term used had to correspond to the original legacy-metadata and then in a "non-filtered" mode where all metadata could be used. After a task had been completed the participant filled in a UEQ questionnaire on that particular task before moving on to the next task. After the entire task list was completed, a short break was taken, after which the participant was interviewed.

Task summary

In preparation of this content retrieval evaluation, the search tasks initially defined were revised based on test panel feedback from the first testing round, and with the availability of more AME metadata in mind. First of all, the tasks were made more realistic with regard to the everyday use of professional media archives. In YLEs media archive, for example, the process of content retrieval is strongly tied to new content producing, the retrieving process often being initiated by a service request made by a content creator who needs to compile material related to the topic of the content being created, most often to a person, subject area, or theme. The tasks were hence refined to better reflect this modus operandi. In particular, the search tasks were modified such that they could be more easily performed using different search modalities without overemphasizing any single auto-generated metadata source. The revised tasks are listed in Table 9. Eventually, tasks 1-4 were such that they could be completed by choosing from several different data modalities, while tasks 5-6 emphasized object recognition that had not been sufficiently tested in the previous round (due to AME metadata from the visual domain that were lacking at that time).

Task	Task Instruction	Evaluation Goal
1	Find a program where Marine Le Pen talks to the press and thanks the French (after her party Front National has won the 2014 European Parliament elections in France).	Basic search for a unique moment in a specific program.
2	Find programs/talk shows where people are discussing the crisis in Ukraine (2 programs from Finland and 2 from France).	Search for a specific topic across two different collections of material.
3	Find a program or clip where Finnish politicians discuss Finns in Syrian camps.	Search for a specific topic in a specific program genre (i.e., politics).
4	Find a program where Belgian politician Nico Cué talks about minimum wage.	Searches clips in which a given person is talking about a topic.
5	Find a clip of a crowd of young people protesting against climate change in Berlin.	Search for program material that contains a specific type of activity in a specific location.
6	Find scenes with domestic animals appearing.	Search for archive material that contains specific visual objects.

Table 9: Revised content retrieval task summary.

As a baseline test, each task was trialed by YLE experts actively working within the MeMAD project, to ensure that achievable tasks were proposed to the evaluation panel.

7.3 Analysis of user data

7.3.1 User Experience Questionnaire

In evaluating Epic 6.3 functionality, the UEQ questionnaire was used after each of the six search tasks (cf. Table 9) to collect the participants' subjective evaluations of the experience of using Flow and the different types of metadata for search purposes. Figure 11 shows the averages of the seven participants' responses to each of the questions on a scale from -3 to +3. The different bars shown for each question represent averages for the six tasks. Averages between -0.8 and +0.8 (shown with dashed lines) are considered neutral, with values exceeding this limit deemed negative or positive.

Overall, the average UEQ scores are neutral or positive, with some differences appearing between the different tasks. Overall the most positive responses are seen for Task 6, where average scores cross the 0.8 threshold into positive assessment for all of the UEQ adjective pairs. All tasks reach a positive average for exciting, and positive averages are also seen for pleasant (Tasks 4, 5, 6), enjoyable (Tasks 1, 4, 6), creative (Tasks 1, 5), motivating (Task 5, 6), relaxed (Tasks 4, 6), fast (Tasks 4, 6), efficient (Tasks 4, 6), fun (Tasks 4, 6), and practical (Tasks 1, 6). Tasks 2 and 3 appear to have caused a less positive response for the participants overall. Average scores for these tasks remain neutral or tend toward negative, although none of them cross the -0.8 threshold into clearly negative assessment. The cases where slightly negative tendencies are observed suggest that the participant found the use of metadata or its quality in Task 2 slightly slower, more inefficient and impractical, and Task 3 also appears to be characterized as somewhat difficult and unpleasant. For Task 1, the experience appears mixed, in that the scores for unpleasant and laborious tend toward negative, but at the same time many other adjective pairs characterize the experience positively.

In addition to the UEQ adjective pairs, the participants were asked to evaluate the quality of different metadata types. Depending on the focus of each task, not all metadata types were relevant for all tasks. In cases where a specific metadata type was not useful to a specific task, the bar representing that task is excluded for the question regarding that metadata. For example, speech transcripts were not used in Task 6, and no assessment is therefore given for ASR in that task. In Task 3, four of the participants opted to use ASR output only, while the three others used both ASR and NER results. In this task, the average score for NER is therefore calculated based on the three participants who used it.

It was observed during initial data processing that some participants had confused some of the metadata types when filling their responses in the questionnaire. This concerned particularly named entity recognition and object/scene classification, and was evident in cases where participants gave a rating to NER in a task where they in fact had object/scene classification and not NER, or vice versa. Prior to calculating the average scores, such cases where it was deemed clear the participant had assessed the wrong metadata type were corrected to reflect the metadata type actually used. These cases could be identified based on the participant’s activity and comments recorded during each task. If the participant had assessed a metadata type that was not used in a specific task, but had also given a score for the metadata type(s) used, such additional scores were excluded.

For the different metadata types, average assessments are also generally positive. ASR was assessed in Tasks 1, 2, 3 and 4, and the average scores are positive in all cases except with a slightly more neutral average in Task 2. Machine translation also receives positive assessments in Tasks 1 and 3, and the average in Task 2 is more neutral although tending toward positive. Named entity recognition was assessed in Tasks 2 and 3, with positive averages in both cases. In Task 3, the average score is based on assessments from three participants who used NER in this task while the other four used only ASR results. Face recognition was used in Tasks 1, 4 and 5, and assessed very positively in each case. Finally, object/scene classification was used in Tasks 5 and 6, with neutral average scores in both cases although tending toward positive.

Summary of comments in the post-task questionnaire

The post-task questionnaires also contained three open questions where the participants were asked to comment on the search experience and quality of the metadata. They were also asked whether they thought some other type of metadata would have been useful for the given task. The comments regarding the search experience and metadata quality were classified into positive and negative comments or in some cases neutral/mixed comments that addressed both positive and negative sides.

Regarding the search experience, the participants made a total of 20 positive, 14 negative and 8 neutral/mixed comments. The most positive comments were made about Task 4 (6 comments), followed by Task 6 (4 comments), Tasks 2, 3 and 4 (3 comments each). Only 1 positive comment was made about Task 1. The most positive comments noted that using the metadata for searching was easy (total 8 comments; Task 2: 3, Task 4: 2, Tasks 3, 5, 6: 1 comment each). The metadata was also described as making the search more efficient (total 5 comments; Task 4: 2, Tasks 1, 3, 6: 1 comment each) and being overall useful (total 2 comments, both about Task 4). The remaining 4 positive comments simply characterized the search experience as "ok". In contrast, negative comments characterized the search experience as difficult (total 6 comments; Task 2: 2, Tasks 1, 3, 4, 5: 1 comment each) or noted that the search brought up no results or wrong results (total 4 comments; Task 3: 2, Task 5: 2). The experience was also described as frustrating (total 3 comments; Tasks 1, 2, 3: 1 comment each) and tiring (1 comment about Task 2). In the mixed comments, the participants noted that searching with the metadata worked sufficiently but that it had some problems or some further development would be needed (total 4 comments; Task 6: 3, Task 1: 1). The remaining comments were more general, noting that the use of metadata took some practice (total 3 comments; Task 1), and that the metadata was fine, but that legacy metadata would already have been sufficient (1 comment about Task 5) to complete the given task.

The metadata quality was generally characterized positively, with 30 positive, 2 negative and 9 neutral/mixed comments. The positive comments mostly described the metadata quality as good, or even "very good", without more specific reference to the type or features of the metadata. The comments were most common about Task 1 (6 comments) followed by Task 4 (5 comments), Tasks 2 and 3 (4 comments each), Task 6 (3 comments) and finally Task 5 (2 comments). Some comments referred to specific metadata types: face recognition was mentioned positively twice (both in Task 5), as was object/scene classification (Tasks 1 and 6 one comment each). One mention was made about both ASR (Task 3) and MT (Task 4). The negative comments were made about Tasks 2 and 5; in both cases the participant simply noted that the quality was not as good as they hoped but did not specify a metadata type. In the mixed comments, the participants stated that the metadata was useful but had some errors or other problems (total 3 comments; Tasks 4, 5, 6: 1 comment each). A specific case was mentioned by a participant noting that in Task 4, the ASR output did not recognize the name “Nico Cué”, but the face recognition results, and hence compensated for audio modality’s lack for a match. One comment about Task 3 noted that the participant had to change the query they used, but after that the search worked. The remaining 5 neutral comments stated that the participant would need more experience to properly assess the metadata.

When asked what other metadata would have been useful, the participants gave varied answers. The most commonly mentioned additional metadata involved more precise topic identification (total 6 comments; Task 3: 3, Task 2: 2, Task 4: 1), followed by some language specific data or additional translations (total 5 comments; Task 2: 2, Tasks 1, 5, 6: 1 comment each). Other types that were mentioned once included content description (Task 1), metadata about the duration of a specific portion (Task 6), metadata about program genres (Task 2) and location data (Task 5). Other comments by the participants noted that they did not need additional type of data but more accurate tags overall (2 comments; Tasks 2, 4: 1 comment each) or more advanced object recognition specifically (1 comment about Task 5). Finally, some participants noted that no other metadata would be needed (1 comment about Task 3) or that they did not know what other metadata might have been useful (total 3 comments about Task 1).

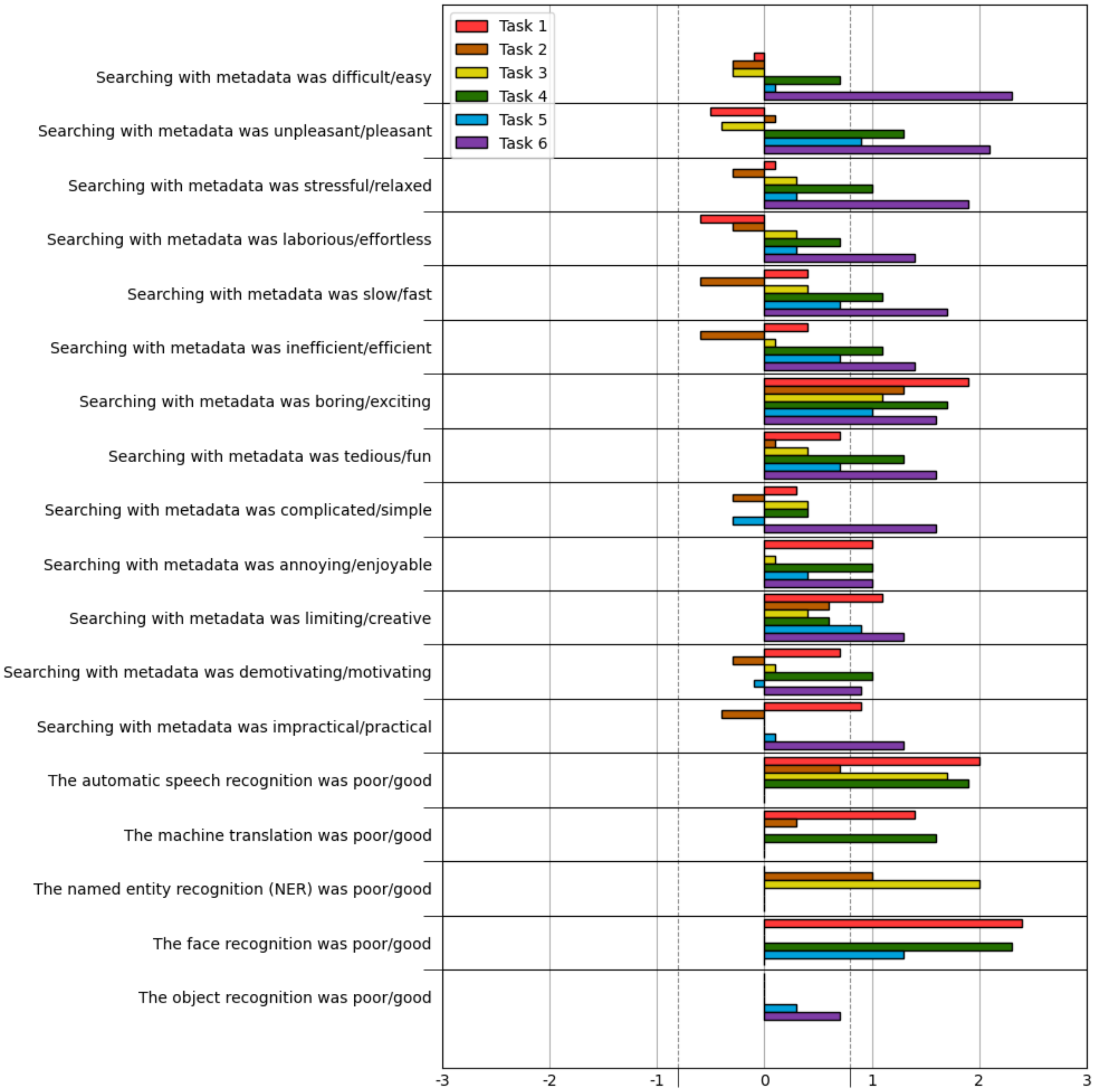


Figure 11: Average UEQ assessment scores for the participants when searching for content in the MemAD prototype platform.

7.3.2 Feedback from interviews

After the completion of the tasks, brief semi-structured interviews were carried out and recorded to collect more detailed feedback from the participants. The following questions and structure were used:

1. What is your overall feeling about the search tasks? (Why is that?)
2. Was there anything positive/negative about the tasks? (Based on the first answer; if their feelings are negative, ask about anything positive.)
3. What features in the metadata impacted the search tasks the most?
4. Did you notice any differences in the metadata quality?
5. How did the use of automatically generated metadata impact your own search process?
6. Could you imagine using this kind of metadata in your work? How should it be improved?

As in the previous evaluation round described in D6.6, the interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements. All users from the test panel participated in the interviews (i.e., four participants from YLE and three participants from Finnish National Audiovisual Institute), which is two participants more than in the previous evaluation round.

Overall, all participants except one expressed a positive feeling about the search tasks and found the experiment interesting. One participant found the experience frustrating because their benchmark is traditional metadata and it was deemed that a highly developed search system could not be matched with the presented automations. Especially at the beginning of the tests, most participants had difficulties with the tool which made it harder to evaluate the quality of the metadata as well as to assess whether it impacted their own search process. Analyzing and verbalizing technical functionalities was also perceived as difficult. One participant regarded the session as stressful due to its length, the amount of new things to absorb and problems encountered in the search process.

Most participants regarded ASR and face recognition as useful and well-functioning in the search. One participant highlighted the usefulness of ASR transcripts in describing factual and discussion programs and another the importance of what a person has said for the retrieval. Two participants mentioned dialects as a potential problem or limitation for the speech recognition, and one participant acknowledged its improvement, stating that the result was now understandable although still partly rubbish. Regarding face recognition, issues of privacy, reliability and its limitation to newer material were mentioned. Only one participant deemed the face recognition negatively, stating that it was primitive and of no use. On the positive side was also MT outputs, even though only four out of seven participants mentioned it. Two participants said the metadata complemented one another well.

On the negative side, the participants referred to technical problems which slowed down the searching. One participant regarded the searching as complicated. Also, finding suitable search terms was deemed difficult: for example, in the test, general terms such as “cattle” were used instead of specific terms like “cow” which are used in the current system as literal object classifications. The search should be user-friendlier, so that the user does not have to think about punctuation or special characters. Support for the Finnish language is important so that the system also retrieves inflected forms. Two participants mentioned the demonstrated object/scene classification which they regarded as a good functionality with new possibilities, if it works well.

Three participants feared that the search result could be too large and contain much irrelevant information. Effective filtering is needed, as well as more possibilities to narrow down the search, for example with regard to time (e.g. the length of the clip, starting from). One participant emphasized the relevance of the working process: being able to find material is only one aspect, another one is the usefulness of the footage that is found, that is, whether there are rights to reuse it and what the picture quality is. Regarding the “tags” being offered by the system, some improvements were suggested: Their number could be more flexibly increased so that they would not exclude anything but find the relevant content. They should be commonly used terms. Further wishes concerned the flexibility of program classification, the search with images and the visibility of images in the search result. One participant hoped for a machine that could retrieve something a human does not typically describe, such as facial expressions, emotional reactions, a tracksuit of a certain brand, or a bicycle of certain color.
All participants could imagine using this kind of metadata in their future work; one, however, with the prerequisite that the automation advances. Two participants considered the system to be complementary to the current one. One participant estimated that the entire profession is going to change; the system requires a new kind of thinking and learning away from the old can be hard, but it will make the data better and more exhaustively utilized.

In comparison with the previous evaluation round (reported in D6.6), we observe some improvement of the system. While problems related to using the system persisted, which had an effect on how the tasks were experienced, some functionalities (ASR, face recognition and MT results) seemed to work reasonably well. The functionalities of automatic object/scene classification and face recognition were new features in this evaluation round, so there was no improvement to measure in this case. Issues for improvement raised by the participants remained the same to a large extent: effective filtering, terminological mismatch between the machine and the human, and the quality of the material that is being searched and found. New, interesting issues regarding object recognition and image-based searching were raised. The participants remained cautious about the automation of metadata and the work process in general, but the feasibility of the technologies seem to have increased: the participants in this round could imagine using this type of metadata in their future work, as opposed to the participants in the previous evaluation who were merely interested in using the tools.

7.4 Discussion of the user feedback and data gathered

Thanks to the auto-generated metadata, the participants had more diverse search methods at their disposal than before. This was mainly seen as a positive feature but sometimes caused extra confusion. In some cases, it clearly depended on which search method was chosen to start with whether a given search task was eventually found easy or difficult.

7.4.1 Observations on AME metadata and their new search methods

As a new feature in the test setup, the participants were allowed to complete all search tasks using Finnish as their search language (except for the metadata tags created from object recognition which were not translated to Finnish).

The usability of MT and ASR collected positive reactions. As the evaluation test progressed, the participants quickly discovered how transcriptions derived from spoken speech could be utilised in the retrieving process. They became soon accustomed to the fact that when a search task dealt with something that someone is known to have said or a topic that someone or some are known to have talked about, it was worth starting a search by looking for results from the automatically generated transcriptions. This was mostly seen as an impressive feature since search tasks were partly related to French-spoken material and none of the participants spoke French fluently.

It was also seen positive that the search process wasn't any more depending on predefined program categorisation and manually added name and topic-tags as in the case in contemporary archive systems. Though it was surprising for the participants how using MT and ASR affected the search process since now they had to focus on words actually spoken instead of trying to figure how a colleague would have recorded the moment in the program's description.

The fact that the participants were allowed to use Finnish as the search language sometimes made them forget the original language used in the material to be searched. This was noticed when answering the questionnaire for each search task. It was sometimes unclear for the participants to perceive whether they had solved the task with machine-translated transcript metadata or just literally transcribed speech from the ASR process. This was noted during the evaluation sessions, and it manifested as well on the collected questionnaire data.

The overall usefulness of the NER disambiguations remained relatively unclear to participants. This could be because MT and ASR produced responses that largely overlapped with the NER results, as the named entities are detected from those sources. As such, the availability of NER outcomes did not make a large perceivable difference to the search panel.

During the test sessions, the reactions on face recognition and visual object/scene classification were mainly positive. Tasks 1 and 4 were such that face recognition played a key role, and 5 and 6 concerned visual object detection. Face recognition was mainly seen as an impressive feature, but visual object detection caused a slightly contradictory reception. This was due to the fact that when searching for domestic animals (Search task 6), the cow was practically the only animal that appeared in the search results.

We did observe that, although all participants had some level of understanding of how the provided metadata was be generated automatically, it was difficult for participants to remember what any technical term such as ASR, NER meant. It was also noticed that when answering the questionnaire for each search task, it was mostly unclear for the participants to perceive which technology they had used to solve that particular task. This was noted during the evaluation sessions and it manifested as well on the collected questionnaire data where the answers were in some questions conflicted in terms of evaluating the metadata used.

7.4.2 Limecraft Flow the basis for the MeMAD prototype test environment

According to the original plan, the participants were supposed to compare two different search results they would obtain when completing the search tasks with two different metadata configurations, first with the legacy-metadata and then with all metadata. Due to some system limitations in integrating metadata on the Limecraft Flow/MeMAD prototype platform, it was difficult to obtain search results that contained metadata from two different temporally aligned metadata sources. Therefore often a zero result was obtained when trying to combine search terms that were from different metadata sources, unless users knew exactly how to instruct the search system to make certain combinations. This situation had an impact on the proceedings of the tests and forced the test coordinator to guide the participants through the process by making suggestions on what 'tricks' could be used to successfully complete the search task.

Limecraft's MeMAD test platform was viewed with somewhat conflicting feelings. Some of the participants felt that the way how things were presented on the platform was confusing since it represented something different from what they were used to in their daily work. It's not certain whether this affected how the usability of auto-generated metadata has been seen, but in some situations, when the automatically generated metadata was questioned, the test organizer considered it necessary to remind the participant that what is being evaluated is the usability of auto-generated metadata and not the platform that served as a test environment.

The observations from this evaluation on this aspect of the prototype platform have been taken into account for future version of the platform. We discuss potential solutions and designs for remediating these shortcomings in Sections 6.3.2 and 6.3.3 of D6.8.

8 Evaluation of Epic 6.11: Intra- and interlingual subtitling

In this section, we discuss the results of three new evaluation studies concerning subtitling with MeMAD prototype software.

The intralingual and interlingual subtitling trials executed in the course of 2019 were repeated with improved MeMAD tools in the course of 2020, and an additional evaluation was added for 2020, namely the evaluation of consumer reception of automated subtitling. For the project, it was important to address both aspects in this final evaluation round for this subtitling epic. First of all, we wished to gauge how the technology developed in the project improved the outcome between each iteration of automated subtitling from the perspective of those people conventionally in charge of the gestation of subtitles: the professional subtitlers. With these improved technologies at our disposal, it was then time to evaluate the usefulness of these automated processes from the side of subtitles consumption. Consumers typically have no deeper understanding of the process behind the making of subtitles, but they have expectations on the quality delivered, so it was crucial to measure their response to automatically generated subtitles. The findings from both evaluation tracks allowed us to make proper conclusions about the productivity and impact of the subtitling work in MeMAD.

In a change of structure from last year's evaluation report D6.6, we have split up each evaluation into its own subsection: Section 8.1 on intralingual subtitling, Section 8.2 on interlingual subtitling, and Section 8.3 on the consumer reception of the automated subtitling pipelines.

8.1 Intralingual subtitling – professional post-editing

As with the evaluations in 2019, this third evaluation round investigated an intralingual subtitling workflow where automatic speech recognition (ASR) was combined with manual post-editing. The evaluation involved professional subtitlers at YLE using the Finnish ASR system developed by Lingsoft and automatic segmentation and timecoding in Limecraft's Flow platform. These tools were used both in production for a proof-of-concept period and a set of final evaluations. The participants' subjective user experience and task times were analyzed.

8.1.1 User test setup

In this section, we provide information about what audiovisual material was used in the evaluation, who participated in it, how the experiment data collection was done, and what tasks users were asked to perform as part of the evaluation.

The intralingual subtitling evaluations consisted of two parts:

1. First, a proof-of-concept period was organized in which subtitlers used MeMAD technology in their daily work. Intralingual subtitling was chosen for this proof-of-concept use based on product maturity and positive feedback from the previous round of evaluations. We also discuss this phase of the evaluation Section 6 of our WP7 proof-of-concept feedback report D7.4.

2. The proof-of-concept phase was followed by more controlled final evaluations resembling those run in late 2019. These final evaluations were conducted to gather data more comparable with the evaluation results of the previous round.

8.1.1.1 Material used

During the proof-of-concept period, participants selected the materials based on whatever they were working on at the time of the experiment, using the MeMAD technology as part of their normal workflow. Most of the materials they selected were factual programs, current events programs and documentaries, although there were a few children's programs and entertainment programs as well. Materials consisted of 14 individual program items, ranging from 10 minutes to 42 minutes in length.

For the final evaluations, video clips were selected from the MeMAD datasets in a single genre, EU election debates, one of the two genres used in the previous round of evaluations. EU election debates were selected as the genre based on the feedback from the previous round of evaluations, where participants found ASR to be more useful in that genre than the more casual lifestyle/cultural programs. Selecting the same genre as in the previous round of evaluations also makes it easier to compare the results.

The individual clips were selected so that each clip 1) formed a coherent, self-contained section of the program as a whole; 2) was approximately 3 minutes long. The length and number of clips was limited due to the limited availability of participants for the experiments.

The intralingual subtitling experiments were carried out in Finnish. Transcripts of the original audio were created using the ASR developed by Lingsoft (which is defined in more detail in D2.2). Subtitles for each clip were then generated automatically in Limecraft’s Flow platform, exported into SRT format and uploaded into individual Drive folders for the participants. The participants then downloaded the files and imported them into the native expert subtitling software used at YLE (Wincaps Q4) for post-editing.

8.1.1.2 Participants

The same professional intralingual subtitlers who participated in round 2 of evaluations in late 2019 were recruited as participants. One of the original participants (FFD) could not participate, so in total there were three participants. All of the participants in this round of testing were in-house subtitlers for the project partner YLE and had between 9 and 20 years of professional experience in subtitling. All three participants had used automatic speech recognition for subtitling before in the form of offline respeaking: two had used it occasionally and one had used it frequently. All three participants had also participated in the previous round of MeMAD evaluations.

8.1.1.3 User data collection and tasks

Due to the pandemic situation, the experiments for subtitling data collection were arranged remotely. The proof-of-concept period took place in the summer of 2020, while the final evaluations were carried out in August and September of 2020, after the proof-of-concept period had ended. The subtitling tasks were carried out using the subtitlers' preferred software environment, Wincaps Q4. The subtitlers had access to the internet as well as terminology and other resources normally used in their work.

During the proof-of-concept phase, the participants used ASR in production for some of their subtitling work. Participants were asked to fill out a feedback form after the completion of each task (see Appendix A for list of questions on the form). The feedback form collected information about time spent on the task, as well as user experience and subjective assessments of the ASR output and automatic timecoding.

For the final evaluations, post-task questionnaires similar to the proof-of-concept ones were filled out after each task (see Appendix A for list of questions on the form). A portion of the form was identical to the proof-of-concept feedback form, but there was an additional section for immediate retrospection, and some of the open questions were different. After the completion of the final evaluations, a semi-structured interview was carried out to collect more detailed feedback regarding problems in the workflow and the participants' views on potential improvements, based on their experience with both the proof-of-concept period and the final evaluations (see Appendix A for the interview script).

In the final evaluations, the participants were instructed to produce subtitles that would be acceptable for broadcasting, and to use the resources (e.g. the internet, terminology resources) they normally would for their work, but to not spend excessive time on "polishing" any given wording or on researching information. No explicit time limit was given for each task, rather, the participants were instructed to work at their own pace.

The following tasks were carried out:

1. Subtitling "from scratch" (1 clip). The participants subtitled the clip without ASR output. Timecoding of the subtitles was also done by the participant.
2. Post-editing of subtitles created automatically with Lingsoft ASR output (2 clips). The participants created subtitles using ASR output and automatic timecoding.

To account for potential differences related to the difficulty of content in each clip, the clips were rotated between tasks.

8.1.2 Analysis of user data

This section presents the results from the intralingual ASR subtitle post-editing evaluations. The results are divided into productivity data (task times and edit distance), subjective evaluations (UEQ scores) as well as the participants' feedback in the post-task questionnaires and semi-structured interviews.

8.1.2.1 Productivity data

No productivity data was gathered for the proof-of-concept stage. The programs selected by the participants varied greatly in length and genre, and there was no data to compare them to in terms of productivity. The proof-of-concept phase thus relied more on user experience feedback.

Figure 12 shows the average task times in the final evaluation for each participant when post-editing ASR compared to the task time when subtitling from scratch. It can be seen that the average task times for ASR+PE (overall mean: 34 min) are in fact slower than subtitling from scratch (overall mean: 30 min). The differences are, however, quite small. The task time averages are also higher when compared to data collected in the first round of ASR+PE experiments in 2019. However, comparing the values directly is problematic. Although the tasks are comparable in that the video clips are of the same length and have similar number of subtitles, and the same participants were involved, it is important to note that the setting where they worked was quite different. In 2019, the experiments were carried out in a controlled setting, with task time measured with a keylogger. In 2020, the participants worked independently at home, and tracked task time based on the time when they started and finished the task. They were also asked to note any longer disruptions and to deduct that time from the total. As the situation was not controlled, however, it is not possible to know whether shorter distractions may have occurred that have affected the overall task time. Statements regarding the effect of ASR on task times cannot therefore be made conclusively. While there was no clear productivity gain, it is interesting to note that the participants' comments (see sub-section 8.2.2.3) on the ASR output and PE indicate that they find it useful, in the 2020 experiments even more so than in the 2019 round.

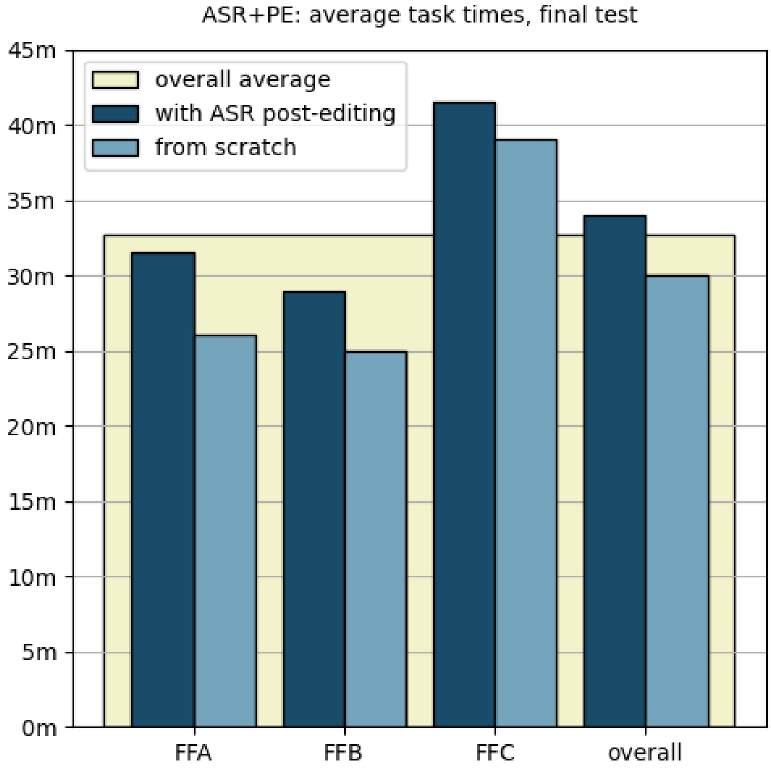


Figure 12: Comparison of task times in intralingual subtitling final evaluation.

To observe the amount of post-editing by the participants, edit distances were calculated for the files post-edited by the participants. Table 10 shows the word error rate (WER) and letter error rate (LER) for each file post-edited by participants FFA, FFB and FFC. The high edit distances indicate that considerable post-editing was done on the ASR output. The LER scores, which calculate the number of changes on the character level, are lower than the word-level scores, which indicates that some of the editing relates to word forms, for example. It should be noted that the level of post-editing does not directly correspond to errors in the ASR output. Due to the technical limits on subtitle frames (number of characters and reading time), condensation and paraphrasing is generally needed especially in programs with fast dialogues such as the election debates used in these tests. For this reason, the final subtitles do not necessarily match the speech exactly. The ASR output also often contains sentences that continue over several subtitle frames, and comparing these to cleanly formatted post-edited versions may also slightly penalize the calculation. Comparison of the unedited ASR output to the post-edited subtitles does not, therefore, directly represent the ASR quality but rather the effort needed to transform automatic ASR output into intralingual subtitles that meet the YLE guidelines for subtitles to be broadcast.

Post-edited	Raw ASR	WER	LER
FFA2 .srt	FFA-2-asr.srt	66.33%	40.91%
FFA3.srt	FFA-3-asr.srt	76.08%	47.55%
FFB-2.srt	FFB-2-asr.srt	88.46%	58.19%
FFB-3.srt	FFB-3-asr.srt	85.31%	56.25%
FFC-2.srt	FFC-2-asr.srt	89.29%	53.61%
FFC-3.srt	FFC-3-asr.srt	81.01%	46.42%

Table 10: The word error rate (WER) and letter error rates (LER) for post-edited files.

8.1.2.2 User Experience Questionnaire

Figure 13 shows the average UEQ scores for the ASR post-editing tasks in the proof-of-concept stage. In this stage, the participants' user experience appears to have been on average neutral to mildly positive. The most positive reactions are seen for the adjective pairs difficult/easy, complicated/simple, stressful/relaxed, unpleasant/pleasant which all cross the 0.8 threshold. The strongest negative response, on the other hand, is for the adjective pair limiting/creative. All others remain in the neutral range. Overall, it appears that the participants tend to characterize ASR post-editing as easy, relaxed, efficient and even pleasant, but limiting and somewhat boring. For questions related to quality of the automatic subtitles, the participants appear to consider the ASR quality high, and timecoding/timing relatively good and easy to correct. Segmentation of the subtitles, however, is characterized as poor. When examining the participants individually, slight differences can be seen in that, on average, the scores given by FFC are more positive while FFB is generally quite neutral and FFA tends more toward negative. However, comparing the participants' scores in the proof-of-concept stage is not straightforward as they subtitled a different number of programs of different types and genres. In the participants' comments below it appears that the type of program (e.g. scripted monologue vs fast unscripted speech) considerably affected their experience.

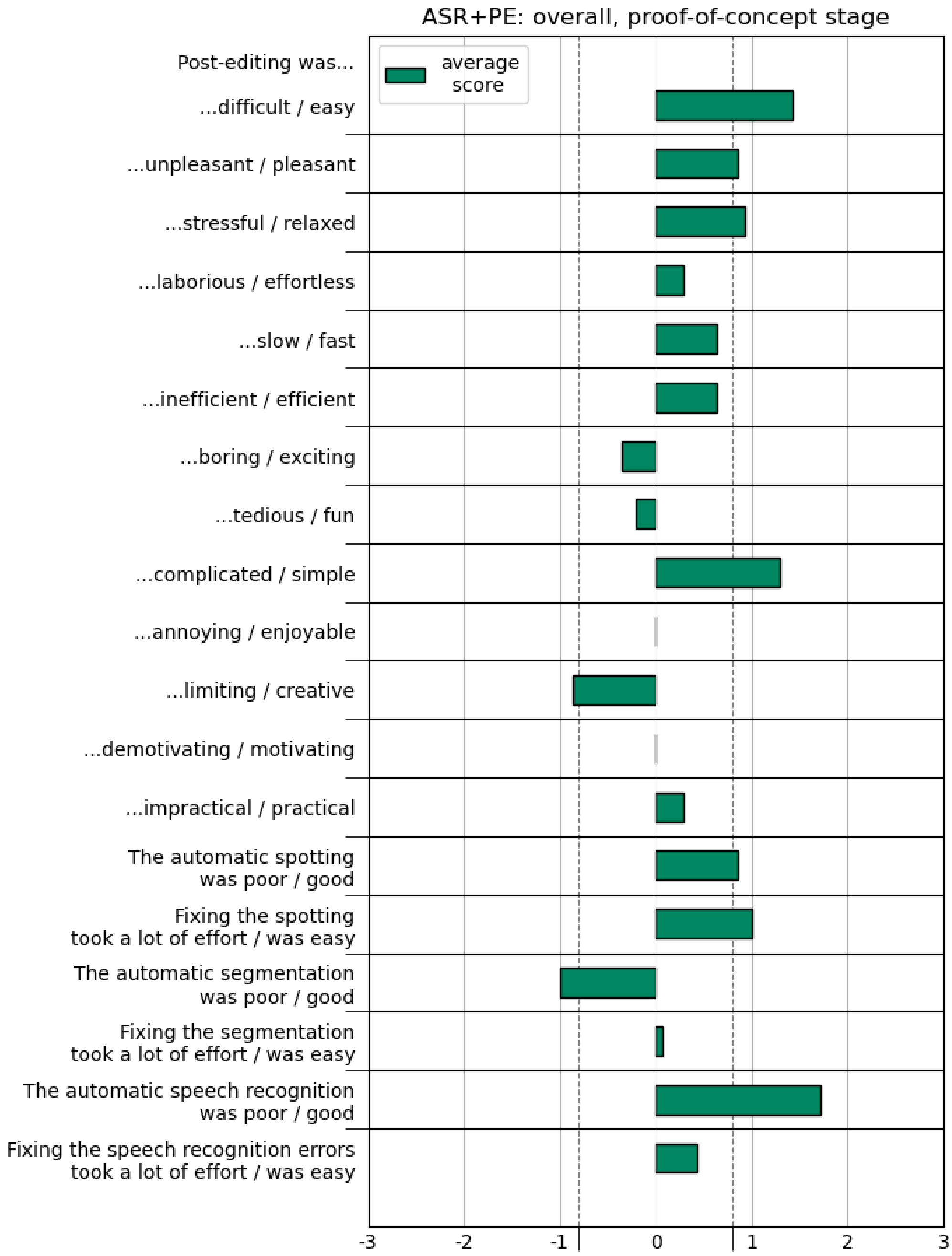


Figure 13: Average UEQ scores for intralingual subtitling, proof-of-concept stage.

Figure 14 shows the average UEQ scores for the intralingual subtitling tasks in the final evaluations. In the final evaluation stage, the participants' assessment of user experience is on average neutral. Although there are more adjective pairs tending toward positive than negative, the only one that crosses the 0.8 threshold to positive assessment is complicated/simple. On the negative side, only limiting/creative crosses the -0.8

threshold, with boring/exciting also approaching this limit. Response to questions about automatic timecoding and segmentation quality are neutral to mildly positive. Differences appeared between the three participants' reactions. FFA expresses the most negative views, with a negative assessment of every adjective pair except complicated/simple. In contrast, FFC expresses very positive assessments for every adjective pair except limiting/creative. Finally, FFB gives a neutral evaluation for all adjective pairs. With regard to timecoding and segmentation quality, FFC also appears to have the most positive opinion, whereas FFA and FFB are more neutral, tending toward negative.

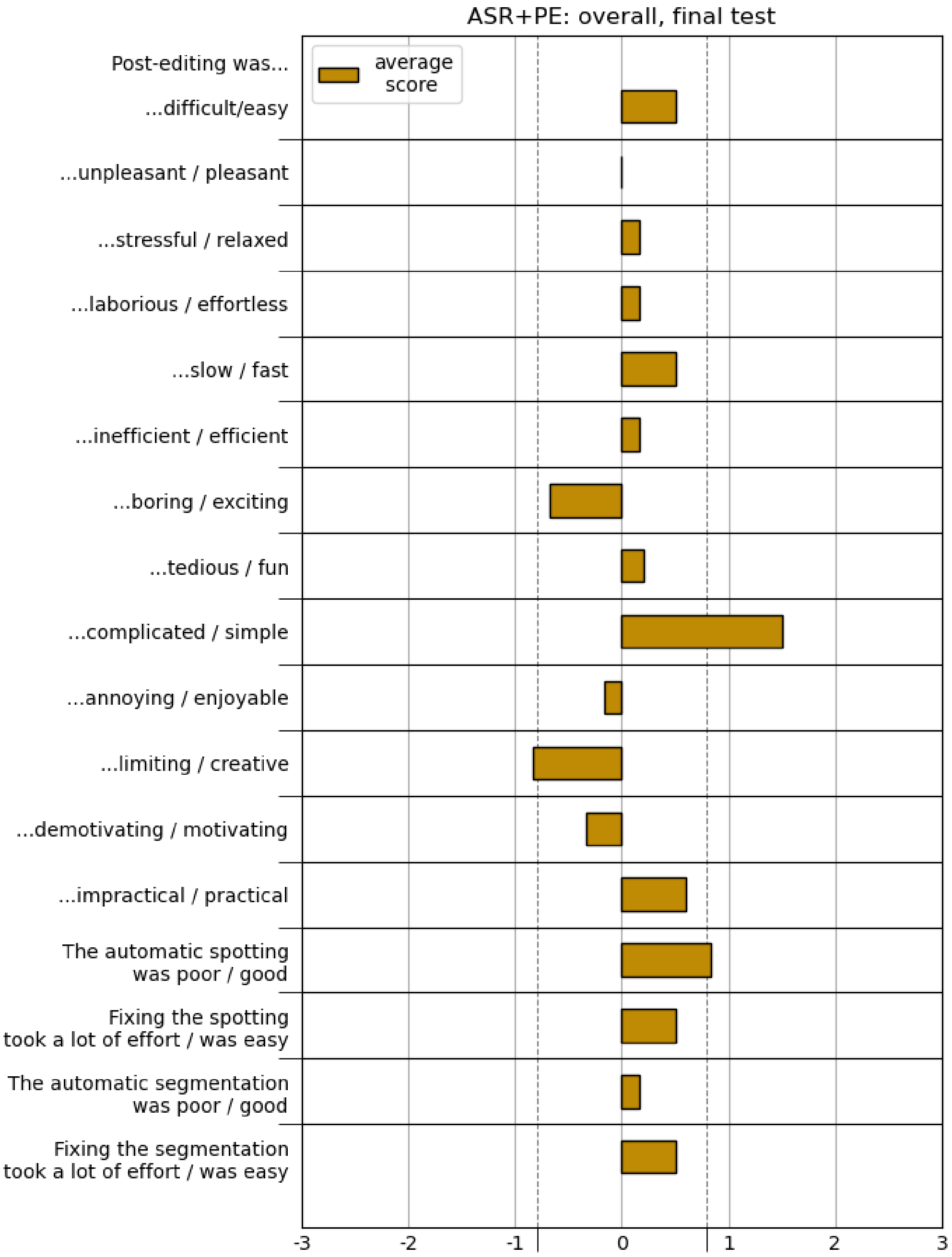


Figure 14: Average UEQ scores for intralingual subtitling, final evaluations.

8.1.2.3 Comments on the ASR post-editing process

For the 14 tasks total in the proof-of-concept stage, most comments regarding their overall experience of the process were mixed or neutral (8 comments). Mostly these comments concerned features of the program and speakers more so than the technology as such. The participants noted that the automatic subtitles were good in parts with only one speaker (e.g. program host) and calm, scripted speech, but struggled with passages with multiple speakers and fast, unscripted dialogues. They also mentioned specific speakers who appeared to have unclear articulation or a foreign accent, for example. Another theme emerging in the mixed comments was that the participants found the ASR output helpful in that it reduced the need for typing but at the same time, looking for errors and the frequent need to delete unnecessary words (repetitions, hesitations etc.) was tiring and annoying. Positive comments (5 in total) noted that the output quality was good, and post-editing was quick and easy. Only one comment was clearly negative; the participant stated that a specific program with dense, fast dialogue required so much condensation that it was annoying to work with.

In the final evaluations, the participants were asked to describe what stages their workflow included in addition to the actual ASR post-editing or creating subtitles from scratch. One of the three participants (FFB) did not specify any workflow steps other than the subtitling phase. Participant FFA described the workflow in the "from scratch" task as using the aligner tool of the subtitling software for automatic timecoding and finalising the subtitles during review in Replay mode, and in the ASR post-editing tasks as "polishing" the subtitles during review in Replay mode, without a separate review afterwards. FFC stated that in addition to subtitling as such, they reviewed the video with edited subtitles, in both the from scratch task and post-editing tasks.

8.1.2.4 Comments about ASR quality

In the proof-of-concept stage, the participants made a total of 30 comments about ASR errors they had encountered in the programs. Mostly the comments concerned general mentions of misrecognized words (7 comments), with more specific comments mentioning incorrect vowel or consonant length (e.g. *niittää* 'reap' instead of *niitä* partitive form of *ne* 'they'; 5 comments), proper names (4 comments) or non-Finnish words (3 comments), and compound words written separately (3 comments). The participants also noted some problems in the presentation of numbers (4 comments) such as showing large numbers as numerals instead of the conventional form (e.g. *750 000 000 000* instead of *750 miljardia* '750 billion') or incorrectly used case endings in numbers. Typographic issues like incorrect capitalization and missing accents in some (foreign) names were also mentioned (3 comments). One of the participants also mentioned cases where words were incorrectly replaced with punctuation (e.g. *myynti.* instead of *myyntipiste* 'salespoint' where the polysemous word *piste* 'point' had been incorrectly interpreted as its other meaning of full stop and replaced with the punctuation character). One participant also mentioned omissions of words. On the other hand, 10 positive comments were made about ASR quality. Mostly these were general comments, noting that the ASR was overall good (5 comments) and there were no recurring errors (3 comments). In two cases, the participant mentioned that proper names, in particular, were usually correct.

In the final evaluations, when asked about errors or issues in the ASR output, the participants mentioned general recognition errors (total 6 comments) affecting proper names (3 comments), other words (e.g. *kansallinen* 'national' instead of *kansainvälinen* 'international'), or differences affecting the morphological form of the word (e.g. the past participle form of *voida* 'can' in the singular *voinut* instead of plural *voineet*) or vowel harmony (*antisemitistisia* instead of *antisemitistisiä* plural partitive of 'antisemitic'). Incorrect capitalisation and compound words were also mentioned (1 comment each). Overall, however, the participants characterised the speech recognition quality as very good (total 7 comments), mentioning specifically correct recognition of (most) proper names (3 comments).

The participants were asked whether the quality of the ASR subtitles differed from the outputs in the previous round of experiments in late 2019. Two of the participants stated that they found the experience and quality similar to the previous year, noting that the quality appeared to be high in both cases. One participant (FFB) stated that in one clip output quality appeared better than in 2019, and for the other expressed being unsure whether there was any difference.

8.1.2.5 Comments about timecoding

In the proof-of-concept stage, the participants made a total of 27 comments about problems with the timecoding of the automatically generated subtitles. The most common issue mentioned (11 comments) was that sentences were split incorrectly, not following linguistic divisions. To some extent, these appear to be related to the issue that periods were missing at the end of an utterance (7 comments), including at speaker changes in some cases (3 comments). Other diarisation errors in recognising speaker changes were also mentioned (3 comments). Other punctuation issues included extra periods in the middle of an utterance (1 comment) and missing commas (1 comment). Timing of the subtitles, as such, was mentioned only once, although some of the participants mentioned that incorrect segmentation led them to adjust also the in or out times. All three participants also made some positive comments about the timecoding, noting that it was overall good.

When asked about issues in the segmentation and timing of subtitles in the final evaluations, the participants commented on incorrect splitting of segments and lines (3 comments) as well as needing to adjust the timing (2 comments), and one participant mentioned that sentence boundaries were not always identified correctly. However, in most of their comments, the participants considered the automatic timecoding to be good (total 8 comments), noting that errors in both the ASR output and in timecoding were not very frequent and were mostly easy to correct.

8.1.2.6 Other comments

In the proof-of-concept feedback form, two of the participants discussed the fact that the ASR subtitles generally required much condensation (total 5 comments). This was connected to the ASR output frequently including hesitations and repetitions - one participant called the ASR "too precise" in this sense. One of the participants (FFC) also mentioned feeling limited by the wording in the output, and finding it difficult to come up with a better, more condensed phrasing. One participant (FFB) noted that the ASR appeared to standardise dialectal expressions, while another (FFC) felt the the output was directing them in a more colloquial direction (e.g. dropped consonants at end of words such as *kuitenki* instead of *kuitenkin* 'although'). Unclear articulation by some speakers (7 comments), fast unscripted speech (4 comments) and background noise (1 comment) were mentioned as features that seemed to cause problems for the ASR. With regard to effort, the participants made comments both that using the ASR made their work easier (4 comments) but also that sometimes correcting took a lot of effort or was tiring (4 comments). They also noted that small errors like one letter differences in case endings were often very difficult to notice and might slip through (4 comments).

In the final evaluations feedback form, the participants commented on various features of the clips themselves and how these affected the post-editing experience more than issues with the ASR output or timecoding, as such. Most commonly the participants commented on the speed and style of speech (10 comments), noting again that fast, unscripted speech and passages with rapid speaker changes led to considerable need for editing to condense (3 comments) and adjust reading speed (1 comment), even if the ASR output was correct. Hesitations and repetitions in the ASR output also led to need for condensation, but one of the participants specifically noted that this type of editing ("picking out the unnecessary words") was very quick and easy. Some specific speakers were also noted to be speaking particularly clearly or unclearly (3 comments). Of the effects of the ASR output, the participants mentioned that it was sometimes difficult to come up with a better, condensed solution after seeing the output, and also that the output affected the register toward a more colloquial style. Related to this, one participant discussed the need to represent the speaker's own style of expression.

8.1.2.7 Feedback from interviews

The interviews with intralingual subtitlers were carried out after they had completed the final evaluations, and the questions touched upon both the proof-of-concept period and observations from the final evaluations. Overall the tone of the interviews was somewhat neutral but tending toward positive: the participants made a total of 11 positive comments, 9 negative comments and 9 neutral or mixed comments about the quality of the automatically generated subtitles or the post-editing experience in general. When asked whether they noticed differences in the quality of the subtitles compared to the previous experiments in 2019, two participants stated that the quality appeared similar - one specifying that the quality had been good in both experiments - and the third did not recall the previous quality well enough to comment.

All participants commented on the ASR output itself being of high quality overall, one of them noting that for some particular content, the output was very close to ready for broadcasting on its own. Like in the post-task questionnaires, they did, however, point out some errors requiring correction. Recurring error types mentioned included incorrect case endings (2 statements), vowel harmony, compound words, proper names and punctuation (1 statement each). While they brought up these issues as recurring, the participants noted that they were often also correct; for example, the participant who mentioned proper names as an issue qualified the statement that mostly proper names appeared to be recognized quite well, but with occasional errors. The point about punctuation errors in the output was further elaborated on by one of the participants stating that this related to the system not recognizing sentence boundaries correctly so sentence ending punctuation sometimes did not appear even at places where the length of a pause signaled a clear sentence break in the opinion of the subtitler. This then led to the automatically generated subtitles to sometimes be segmented incorrectly. This comment was the only mention of timecoding problems made in the interviews. When asked about what improvements would be needed in the future to improve ASR as a subtitling tool, the participants pointed to the same issues as well as overall improving the accuracy of the speech recognition.

When asked whether they would use the automatically generated subtitles in their own work, one of the participants answered yes, because they found the quality to be so high. The other two participants qualified their responses somewhat, stating that they would use ASR for some specific type of content. One of them also noted that it would be important to leave the decision up to the subtitler whether they thought ASR would help with a specific program. According to this participant, sometimes the choice might also not depend just on the genre, but also on the specific situation and whether the subtitler preferred to edit something that was already there or "start with a clean slate".

All of the participants also mentioned variation in terms of the ASR quality and related post-editing effort in different programs they had experimented with in the proof-of-concept or final evaluation. Overall the participants made many mentions of the type of content that was particularly easy to edit and suitable for ASR compared to content that was difficult to edit and less suited for ASR. Content types where ASR quality was best and which were therefore easiest to edit were characterized as those that had slow, calm pace, not more than two speakers, where the speakers were speaking clearly and in a formal, deliberate manner. Content types where the ASR struggled and which therefore caused considerable effort in post-editing included programs with more spontaneous speech, colloquial style and multiple people with interruptions and unclear speech.

Two of the participants also discussed the effects of the ASR output and post-editing on their work. Both addressed the question of effort, noting that sometimes the ASR was helpful because there was not much to correct, but in some other cases with less suitable content quite a lot of editing was still needed. One of the participants (FFB) also talked about how post-editing felt different from regular subtitling from scratch in that it was in some ways less straightforward and needed more mental processing and a different approach. Another participant (FFC) also brought up some concerns about using ASR, specifically that small errors were sometimes difficult to notice, and that post-editing could limit creativity by making it harder to find good solutions for condensing the speech. On the other hand, this participant noted that seeing the ASR output could also have a positive effect by making it easier to start the task. Both FFB and FFC also noted that getting more experience made working with ASR easier. The third participant (FFA) did not discuss any specific effects in detail, simply stating that the experience was interesting and expressing an overall positive view.

8.1.2.8 Final thoughts

While the final evaluations did not show ASR and automatic segmentation and timecoding to improve the productivity of the participants, all participants expressed positive views about both the proof-of-concept stage and the final evaluations. All participants were willing, and in some cases eager, to use this kind of tool in their daily work. Based on product maturity and the feedback from the participants, YLE is planning to use this technology in production.

8.2 Interlingual subtitling – professional post-editing

As was the case in the evaluations in 2019, this third evaluation round investigated an interlingual subtitling workflow where machine translation (MT) was combined with manual post-editing. The evaluation involved professional subtitle translators working in-house or as freelancers for YLE who post-edited MT output generated using either human-created intralingual subtitles or speech-to-text output as the source text. The subtitle MT pipeline was developed by the University of Helsinki. For the speech-to-text cases, ASR, segmentation and timecoding was provided by Lingsoft or Google via Limecraft's Flow platform. The participants' subjective user experience and task times were analyzed.

8.2.1 User test setup

In this section, we provide information about what audiovisual material was used in the evaluation, who participated in it, how the experiment data collection was done, and what tasks users were asked to perform as part of the evaluation.

8.2.1.1 Material used

For the MT subtitle post-editing evaluations, video clips in English, Swedish and Finnish were selected from the MeMAD datasets in a single genre, EU election debates, one of the two genres used in the previous round of evaluations. EU election debates were selected as the genre based on the feedback from the previous round of evaluations, where participants found MT to be more useful in that genre than the more casual lifestyle/cultural programs. Selecting the same genre as in the previous round of evaluations also makes it easier to compare the results.

The individual clips were selected so that each clip 1) formed a coherent, self-contained section of the program as a whole; 2) was approximately 3 minutes long. The length and number of clips was limited due to the limited availability of participants for the experiments.
Each participant post-edited the subtitles of five or six clips, depending on the language pair. English to Finnish, Finnish to English and Finnish to Swedish had six clips each, while Swedish to Finnish had five clips.

In four clips, existing intralingual subtitles were used as the source. The translation pipeline used for these subtitles has been made available through a public repository on Github[7] with pretrained models distributed via Zenodo[8], so that the subtitle translations can be replicated easily and accurately. The pipeline is based on its predecessor from the second year of MeMAD, improved with new translation and preprocessing models trained on larger collections of data. Our subtitle translation pipeline works by first heuristically reconstructing sentences from SRT-formatted subtitles, and converting them to a plain text format with one sentence per line. This conversion process makes

[7] Cf. https://github.com/MeMAD-project/subtitle-translation
[8] Cf. https://zenodo.org/record/4389209#.YBKY4Ogzapo

use of the subalign[9] and OpusTools-perl[10] libraries, both of which we also make publicly available. The translation of the converted sentences was carried out by first preprocessing the source language sentences using general purpose utility scripts from the Moses[11] toolkit [5]. Next, we process them further using our restoration models, which standardize the occurrences of punctuation and letter cases to optimize the sentences for translation. Afterwards, the translation stage receives the normalized sentences, and translates them to the target language. Both the restoration and translation systems are based on the transformer implementation of Marian[12], an open source neural machine translation framework [6]. The former is intralingual translation (from arbitrary to normalized sentences), trained on all available monolingual data from the OPUS collection[13]. The latter involves regular interlingual translation models (from the source to the target language), trained on all available bilingual data from OPUS for each language pair, except for a held-out internal development set sampled from the OpenSubtitles[14] corpus. Once the sentences are translated, we apply general purpose postprocessing scripts from the Moses toolkit, and fit the resulting sentences back into the timed segments of the original intralingual subtitles using the same libraries as in the initial conversion stage. Further details were provided within the Github repository for the subtitle translation pipeline.

In two clips, intralingual subtitles were created with ASR (Google-based for English and Swedish, Lingsoft for Finnish) and the automatic segmentation and timecoding of Limecraft Flow. Translation for these computer-generated subtitles follow largely the same pipeline as described above, except for minor tweaks to the preprocessing stage before restoration, and the postprocessing stage after translation. These stages were both enriched with a few regular expression-based replacement rules (e.g. expanding abbreviations and normalizing punctuation) to address some cases commonly occurring in the ASR-based subtitles that led to repeated translation mistakes. For instance, period characters used as decimal separators (e.g. in “3.14”) or abbreviation markers (e.g. “Mrs.”) have been replaced with special symbols before restoration, then converted back to a period after translation, in order to prevent the neural components from interpreting them as full stops and making translation errors based on that. Another example is that, when translating from Finnish, the abbreviation “n.” (“approx.”) preceding a number has been expanded to the full form “noin” (“approximately”), since the system sometimes did not have access to enough context to know how to translate the single-letter abbreviation correctly.

Swedish ASR quality was found to be particularly problematic. This is probably due to the fact that the speakers in the YLE clips mostly use the variant of Swedish spoken in Finland, which differs from Swedish spoken in Sweden. Because ASR systems for Swedish are mainly trained on data from Sweden (due to relative scarcity of Finland-Swedish data), ASR quality for the variant spoken in Finland is generally lower. Due to

[9] Cf. https://github.com/Helsinki-NLP/subalign
[10] Cf. https://github.com/Helsinki-NLP/OpusTools-perl
[11] Cf. http://www.statmt.org/moses
[12] Cf. https://marian-nmt.github.io
[13] Cf. http://opus.nlpl.eu
[14] Cf. https://www.opensubtitles.com

the frequency of recognition errors, also the subsequent machine translations from the Swedish ASR output into Finnish were deemed to be of such low quality that it was decided to use only one asr+memad clip for experimentation in the sv-fi language pair. The resulting subtitles were uploaded into individual Drive folders for each participant. Participants then downloaded the SRT files and imported them into their subtitling software (Wincaps Q4 or Spot, depending on the participant) for post-editing.

8.2.1.2 Participants

Nine translators participated in the interlingual subtitling experiment, three in the Swedish to Finnish language pair and two in each of the other language pairs (Finnish to Swedish, English to Finnish and Finnish to English). The same translators had also participated in the previous round of evaluations in late 2019, but three of the twelve who took part in the previous round were not available in 2020 due to reasons unrelated to this study. Six of the participants were in-house translators working for the project partner YLE, and three were freelance translators. All participants were professional translators with between 4 and 30 years of professional subtitling experience in the language pair in question. Of the nine participants, only two indicated they had previously used MT specifically for subtitling aside from the 2019 evaluations, although most had used MT for other purposes.

8.2.1.3 User data collection and tasks

Due to the pandemic situation, the experiments for subtitling data collection were arranged remotely. The evaluation period took place in the summer and early fall of 2020. The subtitling tasks were carried out using the subtitlers' preferred software environment, Wincaps Q4 for the in-house subtitlers at YLE and one freelancer, and Spot for two of the freelancers. The participants were asked to record the time spent on each task to the nearest minute and report it in the post-task questionnaire. They were also asked to note any larger disruptions that may have affected the overall task time. Post-task questionnaires were also used to collect subjective assessments of the MT output and user experience after each task (see Appendix B for list of questions on the form). After the completion of the final evaluations, a semi-structured interview was carried out to collect more detailed feedback regarding the workflow and the participants' views on potential improvements, based on their experience with the evaluations (see Appendix B for interview script).

The participants were instructed to produce subtitles that would be acceptable for broadcasting, and to use the resources (e.g. the internet, terminology resources) they normally would for their work, but to not spend excessive time on “polishing” any given wording or on researching information. No explicit time limit was given for each task, rather, the participants were instructed to work at their own pace.

The following tasks were carried out:

1. **intra+memad:** Post-editing of MT output generated with the MeMAD system using intralingual subtitles as source text (2 clips). The participants created

subtitles using MT output and timecoding based on pre-existing intralingual subtitles.

2. **asr+memad:** Post-editing of MT output generated with the MeMAD system using ASR output as source text (1 or 2 clips depending on language pair). The participants created subtitles using MT output and timecoding based on subtitles created with ASR and automatic segmentation and timecoding.
3. **intra+google:** Post-editing of MT output generated with Google Translate using intralingual subtitles as source text (2 clips). The participants created subtitles using MT output and timecoding based on pre-existing intralingual subtitles.

The tasks were ordered so that the participants first completed the intra+memad tasks, followed by asr+memad tasks and finally intra+google tasks. To account for potential differences related to the difficulty of content in each clip, the clips were rotated between tasks, so that each clip was post-edited once with each output by a different participant.

8.2.2 Analysis of user data

This section presents the results from the interlingual MT subtitle post-editing evaluations. The results are divided into productivity data (task times and edit distance), subjective evaluations (UEQ scores) as well as the participants' feedback in the post-task questionnaires and semi-structured interviews.

8.2.2.1 Productivity data

Figure 15 shows the distribution of task times reported by the participants for the different output types. Across all language pairs, average task times for the two cases where intralingual subtitles were translated were nearly equal between intra+memad (mean 34 min, median 37 min) and intra+google (mean 34 min, median 35 min). Task times were slower for asr+memad tasks (mean 46 min, median 46 min). Using ASR output and automatically generated subtitles as the source text for MT appears to slow the participants' work. This is reflected in the more negative commentary regarding the asr+memad output discussed below.

Some differences can be observed between language pairs. On average, the fastest times are seen in the language pair fi-en (mean intra+memad 26 min, intra+google 23 min, asr+memad 39 min), followed by fi-sv (mean intra+memad 28 min, intra+google 27 min, asr+memad 50 min). The slowest average times are seen for sv-fi en (means intra+memad 40 min, intra+google 46 min, asr+memad 58 min). Interestingly, in the language pair en-fi, task times are on average equal for all outputs (means intra+memad 40 min, intra+google 40 min, asr+memad 40 min).

Although overall trends can be seen particularly when comparing the asr+memad outputs to the cases where intralingual subtitles were used as source, it should be noted that task times are also affected by which process stages each participant carried out. In the post-task questionnaire, the participants were also asked which process stages they carried out before or after the actual post-editing. Six out of the nine participants previewed the video with MT subtitles before starting the editing, although only two of

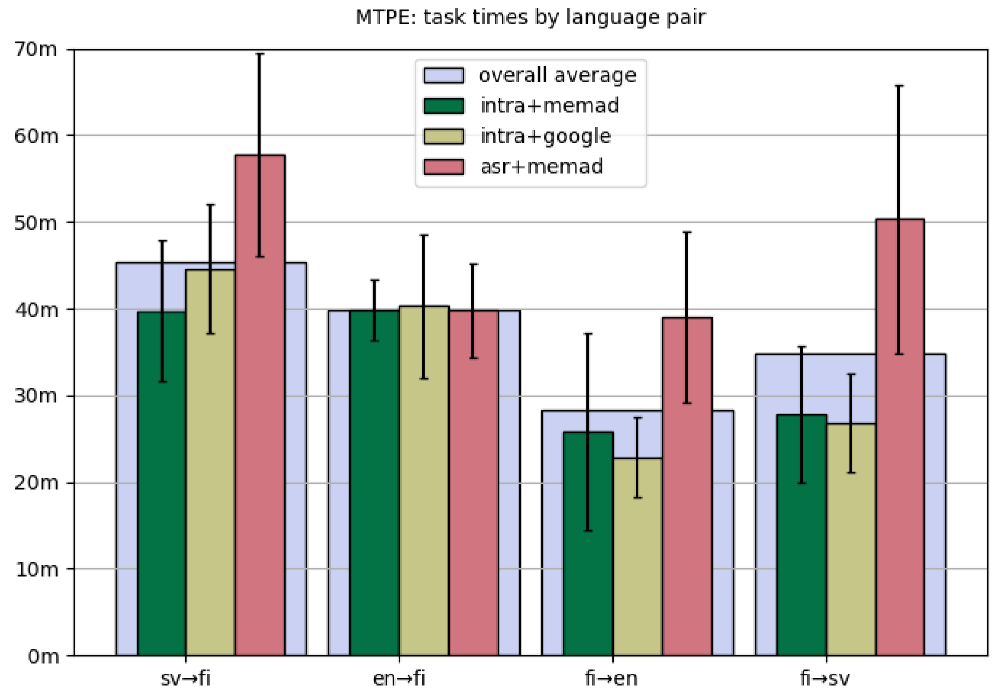


Figure 15: Average task times in the MT subtitle post-editing tasks for each language pair and output type.

them explicitly stated they did this for all clips. One participant previewed some of the clips without MT instead, and another previewed some of the clips both with and without MT. All participants indicated that they reviewed the full clip with edited subtitles after post-editing. Five specified this review stage for all clips, and one for all except one clip, while three only did it for two or three clips. Two of the participants did not specify any additional process stages for some of the clips. In the questionnaire, one participant mentioned wanting to try previewing without MT in order to have an idea of the content before being affected by potentially poor MT quality. Reasons mentioned by others for skipping the final review included lack of time and feeling that the subtitles were "good enough" as they were, but also the opposite (in the case of one of the fi-en asr+memad clips), that the MT had been so poor the participant did not have the motivation to spend time on it.

Some observations can be made with regard to task times in the first user experiment in 2019 and the second round in 2020, although care must be taken when comparing these experiments. In the previous experiments in 2019, the average time for the same participants translating from scratch was on the same level (mean 34 min, median 31 min). However, comparing the values directly is problematic. Although the tasks are comparable in that the video clips are of the same length and have similar number of subtitles, and the same participants were involved, it is important to note that the setting where they worked was quite different. In 2019, the experiments were carried out in a controlled setting, with task time measured with a keylogger. In 2020, the participants worked independently at home, and tracked task time based on the time when they started and finished the task. They were also asked to note any longer disruptions and to deduct that time from the total, and some participants have reported such deductions e.g. due to technical problems. As the situation was not controlled, however, it is not possible to know whether shorter distractions may have occurred that have affected the overall task time. Statements regarding the effect of MT on task times cannot therefore be made conclusively. An overall observation can perhaps be made that effects observed in controlled settings may not always be reflected outside of those settings. In general, the task times appear to be in the same range in both evaluation rounds.

Language pair	System	TER	cTER	Δ words	Δ% words	Δ chars	Δ% chars
all	intra+memad	46.9	0.37	−11	−3.5%	−86	−3.4%
all	intra+google	58.1	0.44	−18	−5.5%	−108	−4.7%
all	asr+memad	88.5	0.56	−98	−23.8%	−621	−22.4%
en-fi	intra+memad	71.7	0.46	−39	−12.9%	−338	−13.7%
en-fi	intra+google	78.6	0.51	−41	−13.6%	−296	−12.3%
en-fi	asr+memad	104.5	0.57	−99	−27.4%	−724	−25.3%
fi-en	intra+memad	32.6	0.30	−12	−3.3%	−53	−2.5%
fi-en	intra+google	33.3	0.27	−29	−7.4%	−136	−6.3%
fi-en	asr+memad	65.8	0.52	−102	−21.2%	−592	−21.5%
fi-sv	intra+memad	45.8	0.42	+1	+0.8%	+3	+0.5%
fi-sv	intra+google	54.2	0.49	−2	−0.6%	−10	−0.5%
fi-sv	asr+memad	95.1	0.63	−127	−27.3%	−720	−26.0%
sv-fi	intra+memad	40.7	0.33	−1	−0.2%	−1	+0.1%
sv-fi	intra+google	64.6	0.49	−4	−1.6%	−14	−0.6%
sv-fi	asr+memad	81.0	0.51	−54	−16.6%	−382	−14.8%

Table 11: Average edit distances on word level (TER) and character level (cTER) and the average change in number of words and number of characters in the subtitle MT post-editing experiments.

To observe the amount of post-editing done by participants, edit distances were calculated.

Table 11 shows average edit distance for each language pair (en-fi, fi-en, fi-sv, sv-fi). The TER score shows the edit distance on the word level, calculating the number of changed words as a proportion of the total words in a given subtitle file. The cTER score calculates the difference on character level. The table also shows the average change in number of words and characters contained in the post-edited files compared to machine-translated files. Similar as in the previous evaluation round, the participants found that considerable editing was needed, which is reflected in the relatively high edit distance scores. The lowest edit distance averages are seen for cases where intralingual subtitles were translated with the MeMAD system, followed by intralingual subtitles translated by Google Translate. This indicates that, on average, the participants considered that less post-editing was needed in the files translated with the MeMAD system. Edit distance scores for the cases where ASR output was translated with the MeMAD system are in all language pairs considerably higher than for the cases where intralingual subtitles were used as a source. In the language pair en-fi, the TER score for asr+memad even exceeds 100, which indicates that the translations have been completely rewritten (the number of changes exceeds the number of words). Some differences can also be seen between the language pairs. On average, the edit distances are lowest for fi-en and highest for en-fi, where the amount of editing remains high for all outputs. The other two language pairs are in between; the average edit distance is lower in sv-fi than fi-sv for intra+memad and asr+memad, but higher for intra+google. It should be noted that the edit distance scores do not necessarily indicate MT errors, as such. Rather, the edit distances are at least partially explained by condensation and paraphrasing typical to subtitle translation.

Comparing the number of words and characters shows that the post-edited files contain, on average, fewer words and characters than the machine-translated files. This indicates that the participants have carried out further condensation of the subtitles, which they also commented on in the post-task questionnaires and interviews. The difference is most notable in en-fi, and to a lesser extent in fi-en. This reduction in the number of words appears to be one factor explaining the overall high edit distances for en-fi. In contrast, for sv-fi and fi-sv, the difference in number of words is negligible in the cases where intralingual subtitles were used as the source. For the cases where ASR output was used as the source text, the reduction is much higher, with nearly a quarter of the MT words being deleted on average. This reflects the difference in the source texts and shows the effect of condensation: the ASR output generally contains all words spoken (although some omissions occur), while the intralingual subtitles already involve condensation. However, particularly in en-fi, it appears that even further condensation was needed in translation.

8.2.2.2 User Experience Questionnaire

Figure 16 shows the average UEQ scores for each language pair and output type. Averaged across all language pairs, UEQ scores for both cases of intralingual subtitle translation remain in the neutral range, mostly tending toward negative. The only adjective pairs that show an overall average toward positive are difficult/easy (intra+memad with a slightly higher average than intra+google), complicated/simple (only intra+memad) and stressful/relaxed (only intra+memad, very slightly). In the other adjective pairs, the average for intra+google is more negative in all but limiting/creative, which is also the only adjective pair where the average crosses the -0.8 threshold into negative assessment for both. The asr+memad output, on the other hand, is clearly received negatively by the participants. For all adjective pairs, the averages are below -1.5 indicating a very strong negative assessment. This negative assessment applies overall to all language pairs, although some variation is seen in individual adjectives. The timecoding and segmentation of the asr+memad is also evaluated very negatively. For the cases where intralingual subtitles were used as source, assessment of timecoding is in the neutral range and the participants characterise it as easy to fix. Segmentation is assessed slightly more negatively, and the responses to ease of fixing segmentation is more neutral.

The en-fi translators appear to have overall a more negative impression of the PE experience: their average UEQ scores cross the -0.8 threshold for negative assessments in every case except for the adjective pairs difficult/easy and complicated/simple, where average score for intra+memad is close to 0. On average, scores for intra+google are lower than intra+memad, and in some adjective pairs even lower than asr+memad. For fi-en, scores are in the neutral range for most adjective pairs. The only case with a positive average is difficult/easy for intra+google; slow/fast and complicated/simple also tend toward positive for intra+google. For intra+memad, slow/fast is in fact in the negative range, as is demotivating/motivating. The adjective pair limiting/creative is

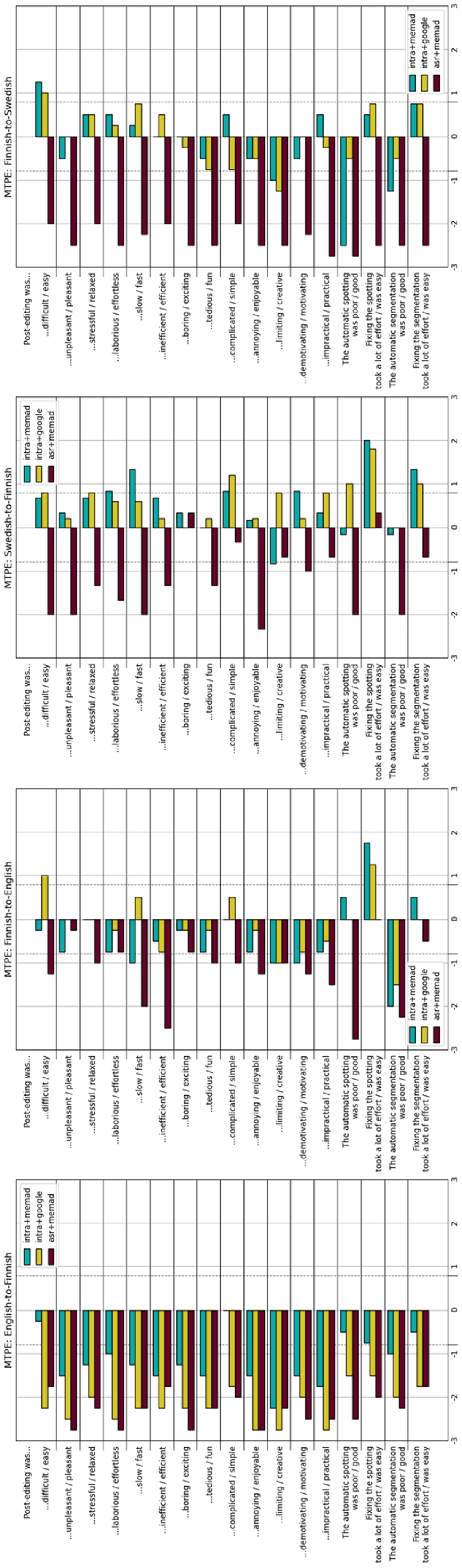


Figure 16: Average UEQ scores for interlingual MT subtitle post-editing by output type for each language pair.

negative for all outputs. Overall averages appear more negative for intra+memad than intra+google, and in one case (unpleasant/pleasant) even more negative than asr+memad. The fi-sv translators show a very negative reaction to the asr+memad outputs, but the two intralingual cases are more neutral. The adjective pair limiting/creative again crosses the -0.8 threshold into negative for all systems, with a slightly lower average for intra+google than intra+memad. The only positive assessment is seen for the adjective pair difficult/easy, with a slightly higher average for intra+memad. Averages tending toward positive can be seen in also some other adjective pairs. The most positive assessments for the two intralingual + MT cases are given by the sv-fi translators. Nearly all adjective pairs are neutral to positive, with positive averages crossing the 0.8 threshold are seen for the adjective pairs complicated/simple (both), slow/fast (intra+memad), laborious/effortless (intra+memad), difficult/easy (intra+google), stressful/relaxed (intra+google), impractical/practical (intra+google). Interestingly, limiting/creative is the only one where intra+memad receives a negative assessment, with an even lower average than asr+memad, while intra+google has a positive average.

8.2.2.3 Comments on MT quality

The participants were first asked to comment on the quality of the MT as such, regardless of the subtitle timecoding. The responses were categorised as positive, negative or mixed/neutral, and specific issues mentioned were further analysed. The responses to questionnaires involving specific output types were analysed separately to identify differences between the outputs.

For the **intra+memad** clips, most participants commented on the clips in mixed or neutral terms, noting both positive and negative features. Clearly positive responses were given by one sv-fi participant and one en-fi participant, while clearly negative responses were made by both fi-sv translators. In the other language pairs, one participant characterized one clip negatively in each case, while all other comments were mixed/neutral. In total, 36 comments referring to specific MT errors were identified in the questionnaire responses. Most comments referred to mistranslations (15 comments), including mistranslated proper names (5 comments) and specific terms (4 comments). Negative comments on fluency were nearly as common (13 comments), including comments on overly literal translations or source language interference (5 comments). Other comments on errors involved omissions (4 comments; all in the language pair sv-fi), missing punctuation (1 comment) and general characterizations of the MT as poor (3 comments). On the other hand, the participants also made 14 positive comments on the MT quality. Half of the comments (7) did not specify any issue, rather, both participants in the language pair en-fi characterized the quality of both clips as overall good (4 comments), and similar comments were made by one participant in each of the other language pairs (3 comments). In the language pairs fi-en and sv-fi participants considered the translations fluent (5 comments), while good accuracy and correctly translated proper names in particular were mentioned by one sv-fi participant (2 comments).

For the **intra+google** clips, comments were also mostly mixed/neutral. Only one clearly positive response was given by one fi-en participant, while clearly negative responses were given by one en-fi participant on both clips, as well as one fi-sv participant and two out of three sv-fi participants. All other comments were mixed/neutral. In total, 37 comments referring to specific MT errors were identified. Most comments referred to poor fluency (17 comments), most often overly literal translations or interference (11 comments). Mentions of mistranslations were also common (13 comments), including mistranslations of proper names (4 comments), but participants did not name specific mistranslated terms in these clips. Typographical issues like extra spaces around punctuation, incorrect capitalization and missing accents in proper names were also mentioned (4 comments; all sv-fi). Some comments did not mention a specific issue, only characterizing the MT as generally poor (2 comments) or unintelligible (1 comment). In contrast, the participants also made 10 positive comments about MT quality. Specific issues included specific useful terms (3 comments), general accuracy (1 comment) or fluency (2 comments), with the other 4 comments being general positive mentions of MT quality.

For the **asr+memad** clips, nearly all responses were negative, characterizing the overall MT quality as bad. No clearly positive responses were given by any of the participants, and only three (one en-fi, fi-en and sv-fi participant) were categorized as mixed, where the participant brought up some positive features in addition to negatives. In total, 27 comments referring to specific MT errors were identified. The answers related to these outputs tended to be less specific than in the other cases, with over half of the comments broadly characterizing the MT as generally poor (9 comments) or unintelligible (6 comments). Specific issues involved mistranslations (7 comments), poor fluency (4 comments, including 2 on overly literal translations) and omissions (1 comment). In the 3 mixed comments, the positives mentioned by the participants were that the translation of one clip seemed "okay" and that some of the terminology was useful.

8.2.2.4 Comments on timecoding quality

Overall, the participants' comments indicate that some of the problems with subtitle segmentation, alignment and timing had improved in the case of subtitles produced using human-generated intralingual subtitles. However, the participants were still not fully satisfied with the alignment and segmentation. Subtitles produced from the ASR output with automatic timecoding rules were evaluated even more negatively.

For **intra+memad and intra+google**, the participants still noted problems with both the segmentation of the subtitles and the timing, regardless of the MT system used to produce the translations themselves. The numbers of negative comments were similar regarding both the timing of subtitles (intra+memad: 11 comments, intra+google: 10 comments) and the segmentation (intra+memad: 12 comments, intra+google: 9 comments). On the other hand, the participants also made some positive comments about the timecoding, more commonly regarding timing (intra+memad: 6 comments, intra+google: 2 comments) than segmentation (intra+memad: 2 comments, intra+google: 3 comments). Additionally, mixed comments (intra+memad: 8 comments, intra+google: 9 comments) indicated that the participant noticed errors but stated that they were easy to correct and limited to only specific frames, after which the timecoding "got back on track".

Comments regarding the **asr+memad** output and automatic timecoding rules of the Flow platform using this fully automated input were generally negative. Across all clips in all language pairs, only one (mildly) positive comment was made, with the participant noting that the timing mostly followed the speaker turns so that it could be used as a template. All other responses regarding timecoding were negative. Comments about timing or alignment problems (14 comments) were slightly more common than issues with segment splitting (10 comments). A common criticism was that the ASR and MT was "too literal" and "tried to include everything", without condensing the message in the way a human would. This then led to the clip containing too many subtitles with too short screen times and problems with text and speech synchrony.

With regard to timing, the subtitles mostly appeared to be delayed, although some cases of the subtitle appearing too early were also mentioned. Sometimes this was caused by alignment issues, where a subtitle segment (or part of it) was placed in an incorrect subtitle frame. These appear to be due to either overly long translation in one frame, which then causes parts of the segment to get "pushed" into the next one, but sometimes also by missing content (parts of segments or even full segments) being omitted in the MT output. Particularly problematic were the cases where misalignment caused a subtitle with one person's utterance to appear during another speaker’s turn. Timing issues also have a connection to segmentation, with participants noting that sometimes the split between subtitle frames comes at an incorrect point in the segment. On the other hand, some participants explicitly noted that these were less common than cases of incorrect line splits within a subtitle frame. They commented that the current way of splitting lines and also subtitle frames does not always follow linguistic units the way a human subtitler would. However, some of the comments regarding timing of subtitles do not in fact relate to problems with the alignment of the MT. Rather, the participants stated they had to correct the in and out times of the subtitle frames, or gaps between subtitle frames, which had been set by the intralingual subtitlers. This suggests that the same timing did not work in the interlingual situation, or may even point to individual differences between subtitlers.

For each clip, the participants were asked to comment on how the quality of the MT subtitles appeared compared to the previous year's clips. The responses varied depending on the output. For **intra+memad** clips, responses were generally favorable, with most participants (all fi-en and sv-fi, one en-fi) stating that the quality of both clips was better than the outputs in the previous experiment. Both fi-sv translators and one en-fi characterized the quality of one clip as better and the other as similar compared to 2019. For **intra+google** the responses were more mixed. Most participants considered the quality to be similar as in 2019 or stated they were not sure. Both fi-en participants and one fi-sv and sv-fi participants found the quality of one clip better, while one en-fi participant considered the quality of one clip worse than in 2019. For **asr+memad** the responses were again generally negative. It should be noted that the previous interlingual evaluations in 2019 investigated only the use of human-created intralingual subtitles as source text for MT, not ASR output. Since the ASR+MT outputs were overall observed to be of lower quality than MT of intralingual subtitles, it is not surprising that most participants either stated that the quality of the asr+memad clips was worse compared to the 2019 clips or generally indicated that the quality of the asr+memad clips was poor. Only one sv-fi participant commented that the **asr+memad** output was possibly better than in 2019. Both fi-sv participants and one sv-fi characterized at least one clip as similar to 2019. One fi-en participant did not answer this question at all for one clip but this participant's other comments indicated they found the quality too poor to work with.

8.2.2.5 Feedback from interviews

In the semi-structured interviews conducted after all tasks had been completed, the participants discussed their overall impressions, and all noted that quality varied in the clips they post-edited. Only one participant brought up differences in timecoding/segmentation, stating that the automatic timecoding was completely unusable in one clip (identified as one of the asr+memad clips based on the participant's reference to the clip number). The other participants referred to varying MT quality. The participants themselves were not aware of how the output for each clip was produced, but based on references to specific clip numbers which the participant recalled as particularly bad or comparison of the interview comments to the questionnaire responses, comments about specific clips with particularly poor quality appear to concern mostly the asr+memad output. The three sv-fi translators made clear references (through clip identifier or specific content) to the asr+memad clips as being worse than others, while one of the fi-en translators noted that the first clips (intra+memad) appeared the best and then the quality decreased. The other participants did not identify specific clips, but based on the questionnaire responses, comments about lower quality likely relate to the asr+memad clips. One participant also mentioned that two of the subtitle files (translated by Google) had contained some character coding problems.

The participants made a total of 18 negative comments that related to MT quality. The most common issue brought up related to accuracy/mistranslations (5 comments) and specifically mistranslated proper names (4 comments). In addition translated names (e.g. *Antti slope* for *Antti Rinne*), one participant mentioned the spelling of a name had changed in MT (*Margrethe* had become *Margareta*). Other mentions of MT quality issues included grammar errors (2 comments) and awkward or unidiomatic language (2 comments). One participant also mentioned typographical issues, specifically mentioning that the MT system did not correctly handle the use of colons in Finnish to add case endings to acronyms or numbers. Four negative comments also referred to MT quality without naming a specific issue. On the other hand, seven out of nine participants also made a total of 12 positive comments about the MT quality. The most common positive characterization was that the translations were fluent and idiomatic, although two of the participants qualified these statements by noting that the fluency could be misleading. In contrast, one of the participants stated that some of the translations were a bit "clumsy" but still useful. Two participants brought up terminology as being helpful, although one of them noted that it was not always entirely reliable. Four comments referred to good quality without specifying an issue, although also here two of the participants qualified that only some of the clips had good translations.

Timecoding (timing and segmentation) of the subtitles still appeared to be problematic, as the participants made a total of 20 negative comments about the timecoding quality. The most common issue was the subtitles being out of sync with the speech (10 comments), followed by incorrect segment splits (7 comments), and too short reading time in some clips (3 comments). One participant also mentioned the splitting of rows within a subtitle frame as a problem. Although the participants were still not completely satisfied with the timecoding, they did note improvements compared to the previous year, at least in the best clips. Only one of the participants, however, specifically characterized overall timecoding quality as good this year.

Similar as in the post-task questionnaires, the participants mostly deemed the machine translated subtitles better in this round compared to 2019. Both of the fi-en translators and one en-fi stated that the 2020 translations had improved. Two of the sv-fi translators also considered the quality of most clips to be better, but noted that some specific clips appeared to be similar or even worse than last year. The other en-fi translator and both of the fi-sv translators described the quality as overall similar in both rounds, but one fi-sv translator noted that some clips this year had worse quality. The third sv-fi translator stated it was difficult to recall differences. As noted above, the comments about worse quality generally appear to be directed at the asr+memad clips. When discussing the differences, the participants noted that timecoding of the subtitles had improved (4 comments), that overall MT quality was better (2 comments), approximately the same (1 comment) or worse in some clips (1 comment). In some cases, the participants referred to the overall quality of the subtitles without specifying whether they meant MT quality, timecoding issues or both.

The participants also observed issues which they did not consider errors, as such, but rather things that caused problems for the MT system and sometimes made post-editing difficult. The need to paraphrase and condense was discussed by two participants. One stated that the MT could not determine the important parts in what was said, and therefore was not able to condense the text like a human would. Another participant mentioned a specific situation like a speaker listing a number of countries which would not all fit in the subtitles, where the translator needs to decide how to handle that list (choose which ones to list, group the countries in some way or some other solution), which the MT could not do. Two participants also discussed the fact that languages may differ in the way things are ordered, for example, how main vs subordinate clauses are organized. One of them noted that occasionally the MT had in fact reordered a sentence in a way that was fluent and even a good translation for some other purpose but which did not work for subtitling because the way the spoken and written text was then out of sync made the subtitles confusing. Two of the participants also mentioned that post-editing was particularly challenging when the speech was unclear because it was difficult to make out what was actually said and not to get misled by potential ASR and/or MT errors.

When asked about what developments would be needed to make MT a more useful tool for subtitle translators, the participants mostly pointed to the quality issues mentioned previously: improved timecoding and synchronization of subtitles (6 comments), improving idiomaticity and overall translation quality (3 comments), more accurate speech recognition (2 comments), better recognition of proper names (1 comment) and creating more condensed subtitles (1 comment). Some participants also suggested other changes to the tools or processes. Two mentioned that they would find it useful to be able to customize the tool, either for their own use or for specific content by integrating terminology resources and fine-tuning the MT for specific programs. One participant also discussed how it might be useful to modify the process so that instead of giving the translator a pre-segmented MT output, the post-editing would happen before timecoding/segmentation or the translator would do the post-editing and timecoding at the same time. This participant also mentioned that they would like to see more information about where the translation was coming from, for example whether the system had been trained to handle a specific type of content, because this might affect how well they could trust the terminology.

The participants were also asked whether they could see other metadata, for example ASR output or recognition, as useful in their subtitling work. Two participants did not discuss any specific technologies or metadata: one was not familiar with these technologies and the other simply noted that additional information was usually good to have. The others appeared to find ASR output potentially most useful. Four of them mentioned that ASR of the original speech content (with or without timecoding) could be helpful. Two of the participants (one en-fi and one fi-en who also does translation from English) mentioned that they sometimes worked with English templates, and that using ASR could be similar. One fi-sv translator stated that it probably depended on one's working style: this participant did not think ASR would be useful for their own work, but some others might find it helpful. Another sv-fi translator mentioned that ASR could be helpful, but also expressed concern that if the quality was not good, it could even hinder the work. The others did not specifically comment on ASR. Usefulness of face recognition appeared more limited. Only one participant stated that face recognition would be useful, referring specifically to when they needed to create a list of text inserts[15]. Two participants stated that they did not see any use for face recognition in their own work, although one of them later amended that perhaps occasionally it could help find names for people the translator did not recognize. This participant also stated that overall it would be useful to have a tool for identifying and checking the names of people and organizations (i.e. named entity recognition, although the participant did not use that term), as well as terminology resources. Two participants stated that they were not really familiar with this technology and would need to test it in practice before they could really answer, and the others did not directly comment on face recognition.

When asked whether they would consider using MT and post-editing in their work, the participants mostly gave mixed answers. Two stated they could see themselves using MT, although both specified that they would want to decide themselves when and how to use or not use the output. Three participants stated that they would not use the MT at the current quality level, but then went on to state that they would be open to the idea once the quality improved so that it clearly made their work easier. The other four participants gave more mixed answers, stating that they would consider MT useful for some content types or in some specific situations but not for all subtitling work. MT was described as most useful for current affairs or news type programs, particularly for rush jobs. However, one participant cautioned that rush jobs sometimes include high-profile content where errors or poor language slipping through might be particularly damaging. Fiction and programs with very colloquial, figurative and expressive speech were mostly

[15] Note: At YLE, an additional action done by translators, particularly in documentaries, is to prepare lists of translated text graphics to be added to the program (e.g. name graphics "John Smith / title" are replaced with translations, rather than subtitles to translate such elements).

mentioned as types where MT would not be useful, although one participant mentioned it might be suited for some "lighter" lifestyle content where the language was "easy".

The participants actively discussed the potential effects of using MT and post-editing. The most common issue brought up was that while the MT output was fluent and often seemed good at a first glance, sometimes the fluency was hiding considerable errors of meaning. Six out of the nine participants made one of more comments referring to this problem (total 11 comments). Some of them expressed concern that this could lead to errors slipping through particularly if the translator was inexperienced, unused to dealing with MT or in a hurry. Additionally, one participant discussed more broadly that it was difficult to trust the MT. The effect of MT on productivity was also a recurring topic. Three of the participants felt that using MT was more efficient in some situations, while three stated that post-editing seemed like a lot of work and doubted whether it was more efficient than translation from scratch. One participant discussed both sides at length, concluding that although they could see some occasional benefits, the experience did not convince them that post-editing was overall more efficient since the time potentially saved by having useful MT suggestions seemed to be taken up by fixing the timecoding of the subtitles. Four others also made some comment that having the pre-generated subtitles had a negative effect on the way they worked. They discussed how the MT seemed to limit their creativity, made it difficult to think of good paraphrases when condensation was needed, and overall made the work more tedious. More neutral comments were made by three other participants who characterized post-editing as very different from their own work processes but did not find it negative, necessarily. These three noted it would take getting used to and learning new approaches. Finally, three participants discussed their concerns that wider use of MT could have negative impacts on the working conditions of the subtitling field, if assumptions about productivity lead to reducing compensations and tighter deadlines, as well as lower quality of subtitles.

8.2.2.6 Final thoughts

While post-editing machine translated subtitles seems to improve productivity for some translators, this was not true for all participants. Even those whose productivity was improved were unenthusiastic about the process. Many of the participants were open to the idea of using machine translation technology in their work, but largely not in the form it was presented to them. Some participants would prefer to have the machine translation as a separate text file to refer to while working, and in previous evaluations they mentioned the idea of having the machine translation work “more like translation memory software”. Quality remains a concern, in several senses. In terms of technical implementation, improvements are needed both in the machine translation quality and in the segmentation and timecoding. On the other hand, translators are also concerned about the quality of the end product - of errors slipping through, and of the machine translation guiding their choices so that the end product might be “correct”, but not entirely idiomatic or natural. As such, machine translated subtitles were not yet considered to be mature enough to be used in production. Using professional translators in the evaluations provided valuable information to the project, and practitioners’ viewpoints and feedback should be considered an important element in any future research.

8.3 Interlingual subtitling – consumer reception

In the previous round, the evaluations on machine-translated interlingual subtitles focused on the perspective of professional subtitlers who used MT subtitles for post-editing. The evaluation report (cf. D6.6, p. 58) mentioned that it is an open question “what subtitle quality would be good enough in order to be comprehensible and sufficiently readable by consumer (viewer) end users”. In order to assess the quality, usefulness and future potential of MT subtitles, it is therefore necessary to gather data directly from end users. In addition to comprehensibility and readability, these evaluations explored the acceptability of MT subtitles to ordinary viewers. The key questions in the evaluations were:

1. How comprehensible are video clips with raw machine-translated subtitles that have been created with MeMAD tools?
2. How acceptable are raw machine-translated and fully automated subtitles to viewers? What are some key factors that determine subtitle quality and acceptability from viewers’ point of view?
3. How much cognitive load do raw machine-translated subtitles cause for viewers in comparison to professionally produced subtitles?
4. What use contexts do viewers envision for raw machine-translated subtitles?

Viewers’ opinions are unique and subjective, but sufficient empirical data on their authentic perspectives can give a realistic indication of the reception experience. In addition to immediate opinions, viewer research can shed light on attitudes and expectations, which is helpful for determining appropriate messaging on MT subtitles.

8.3.1 Viewer test setup

From the perspective of the audience, the development of the MeMAD tools is still in an exploratory stage where different approaches are tested and their benefits and drawbacks explored. At this stage, it is useful to collect qualitative, open-ended data that allows viewers to express their views without being narrowly constrained to a specific question. That approach has the potential to produce broad-ranging data that will inform further development and help shape more detailed questions for future studies. We were able to combine this broad, exploratory approach with a step towards more specific questioning by designing a three-stage evaluation. First, we conducted two focus group discussions as a way of gauging general attitudes. Then we designed a questionnaire to gather more detailed quantitative data on specific questions based on findings from the focus groups. Finally, we conducted a second set of two focus groups to provide further qualitative data on more well-formed questions based on the first two steps. This design allowed us to gather multiple types of data with a focus on exploratory questions.

The focus groups were conducted with two language pairs: translations from Swedish into Finnish were tested with Finnish-speakers, and translations from Finnish into English were tested with English-speakers. The questionnaire was in English and aimed at native or near-native English speakers, who were shown video clips translated from Finnish into English. The first focus group session was conducted with one video clip for each language, the questionnaire tested viewer reactions to two slightly shorter video clips, and the final focus groups were conducted with two shorter clips each.

To answer the four research questions, they were operationalized into four themes that were used to structure the focus groups and the questionnaire. The themes were 1) comprehension, 2) cognitive load, 3) appreciation of and reactions towards the sample clips, and 4) thoughts on the usefulness of automated subtitling (see Appendix C for both focus group scripts and the questionnaire).

8.3.1.1 Material used

The decision was made to limit most viewer evaluations to fully automated MT subtitles rather than different versions with varied amounts of human involvement. This was done to test a scenario that was deemed most ecologically valid and likely in practice. Especially in a small language like Finnish, one likely use for MT subtitles is to translate audiovisual material that would otherwise not be translated due to lack of resources. For example, YLE would not be able to assign enough resources for human translation to provide English subtitles for their daily Finnish-language news and current affairs programming, even though this gap in translation provision is recognized. In addition, required turnaround times may be so fast that human involvement would be impossible. Therefore, MT could provide a solution. The first round of focus groups and the questionnaire therefore included only fully automatic subtitles. The second round of focus groups also tested a scenario where the ASR output of one video in each group was slightly edited before the MT to correct specific ASR problems that caused difficulty in the MT phase (see below for more details). The purpose of including these edited versions was to gain a deeper understanding of the readability and acceptability of fully automatic subtitles and to explore whether minor improvements would change viewer attitudes. In addition, this experiment will help explore whether it would be useful to consider workflows where the ASR is edited even if human input is not included at the translation stage.

The first focus groups in summer 2020 were shown an approximately five-minute clip of a current affairs program with fully automated MT subtitles. The Finnish group saw the first five minutes of a Swedish-language documentary from the series *Spotlight* discussing quicksilver emissions of a steel factory in Finnish Lapland. The English group saw the first five minutes of a Finnish-language investigative program from the series *MOT* covering corruption related to international sales of armored vehicles by the Finnish defense industry company Patria. The topics were chosen to be fairly challenging, so that viewers would be unlikely to understand much of the source-language narration. The subtitles were produced automatically with the same ASR (Google for Swedish and Lingsoft for Finnish) and subtitle MT pipeline that was used for the interlingual subtitle post-editing evaluations (see Section 8.2.1.1).

The English questionnaire circulated in fall 2020 contained two clips translated from Finnish into English. One was a shortened version (approximately three and a half minutes) of the same *MOT* documentary clip as in the first focus group. The second was a clip (approximately two minutes) from YLE's local news from Eastern Finland on plans for a faster train route from Helsinki to Eastern Finland. Shorter clips were used in the questionnaire to keep the time needed for answering the questions manageable. The subtitles were again generated fully automatically. Because Lingsoft's Finnish ASR system had been updated during late summer 2020 with improvements especially to the speaker diarization, new subtitles were generated also for the shortened documentary clip to take advantage of the quality improvements. The same subtitle MT pipeline was used as in the interlingual subtitle post-editing evaluations (see Section 8.2.1.1), except that a slight modification was made to the processing of the subtitles generated with ASR. The conversion from subtitle segments to sentences uses a different, more simplistic conversion script, splitting sentences strictly from terminal punctuation (periods, exclamation, etc.), which we have found to work better for these subtitles.

In the second round of focus groups in November 2020, the participants viewed two clips. In the English group these were the same clips that were used in the English questionnaire (shortened version of the *MOT* documentary clip and the news clip about train routes). Similarly, the Finnish focus group was shown a shorter version of the *Spotlight* documentary (approximately three and a half minutes), and news clip (approximately two and a half minutes) from YLE's Swedish-language news on a survey about the experiences of Swedish-speaking Finns during the COVID-19 restrictions in spring 2020. In both cases, the subtitles for the news clips were generated fully automatically with ASR and MT. For the documentary clips, we decided to test some minor manual correction of the ASR output before machine translation. The purpose of this "pre-editing[16]" for MT was to test the potential effect of correcting specific problems in the ASR output which caused particular difficulties for the MT. These problems involved incorrect punctuation due to misidentified sentence boundaries and speaker turns, incorrect number formatting and incorrect proper names (e.g. the city name *Torneå* rendered as *tornado*). The pre-editing was limited to these issues because it was determined that resources for more comprehensive manual correction of the ASR output would likely not be available in a practical scenario. Mechanical corrections to punctuation, number format and proper names could be done rapidly and potentially automatised with future developments of ASR, named entity recognition and other processing steps. The documentary clips were determined to be the more plausible candidates for editing, as the turnover times for news clips are likely to be even shorter than the schedules for documentary programs. Both focus groups were first shown the fully automatic news clip and then the documentary clip with light pre-editing of the ASR for MT and participants' responses to the two were compared. The ASR outputs (unedited for the news clips, lightly pre-edited for the documentary clips) were machine translated using the same MT pipeline as in the previous cases. Also in this case, the conversion from ASR subtitles to sentences uses the simplistic conversion based on terminal punctuation only.

8.3.1.2 Participants

Focus groups

Due to the COVID-19 pandemic and restrictions on face-to-face meetings, all focus groups were conducted online via Google Meet. The participants in all focus groups were

[16] The term pre-editing is used here to clarify that the manual corrections were done prior to the MT phase for the purpose of improving the automatic translation, and to distinguish this process from the practice of post-editing ASR output for intralingual subtitling (see 8.2) or post-editing MT output for interlingual subtitling (see 8.3).

selected primarily on the basis of their language skills: they were expected to speak the target language of the subtitles (Finnish or English) as their native language or at a near-native level, i.e. C2 in the Common European Framework[17]. In addition, they were expected to know as little as possible of the source language in the video clip (Swedish or Finnish), a maximum of B1 in the Common European Framework. The assessment of language skills was based on the participants' self-evaluation. The objective of this selection was to ensure that the participants would rely on the subtitles as much as possible and that their perceptions of the subtitles would not be limited by shortcomings in their knowledge of the target language. The participants were also selected to represent a broad range of backgrounds in terms of age, education and occupation in order to provide a variety of perspectives into the discussions.

The first Finnish focus group took place on 15 June 2020, and it consisted of six participants, two men and four women, ranging in age from 20 to 63. The first English focus group took place on 16 June 2020, and it consisted of seven participants, four men and three women, ranging in age from 24 to 44. Three of the participants were native English speakers and four near-native speakers of English. The second Finnish focus group took place on 26 October 2020, and it consisted of seven participants, four men and three women, ranging in age from 22 to 67. However, one participant had technical difficulties, and this individual's participation was very limited, which means that in practice there were six participants. The second English focus group took place on 27 October 2020, and it consisted of seven participants, four men and three women, ranging in age from 23 to 65. Six of the participants were native English speakers and one was a near-native speaker.

Questionnaire

The English online questionnaire was distributed through social media (e.g. project participants' Twitter and Facebook accounts, the Facebook account of YLE's English-language news service) and other online networks (e.g. the mailing list of the Finnish branch of the Erasmus Student Network) with the purpose of reaching a varied and large group of potential respondents. The introductory text stated that the respondents should be native or near-native speakers of English and at least 18 years old, but there were no other restrictions on participation. Some respondents understood Finnish better than the focus group participants, but there were also a large number of respondents who did not live in Finland and were not familiar with Finnish language and culture at all. This diversity in the respondents allowed us to explore the research questions more deeply, as it provided comparisons between individuals with different linguistic backgrounds and other relevant factors. The background factors the respondents were asked to disclose were their age, education, Finnish skills, experience with viewing subtitled programs, and whether the respondent had ever lived in Finland (see Appendix C for the questionnaire).

The respondents were asked whether they lived or had ever lived in Finland to assess whether cultural familiarity may play a role in comprehension. The question originated in the first English-language focus group, where some participants speculated that it

[17] Cf. online at https://europass.cedefop.europa.eu/fi/resources/european-language-levels-cefr.

may have been easier for them to follow the video because they lived in Finland and were used to the sound and structure of the language than it would be for someone who has had no contact with Finnish.

The questionnaire was open from 7 October to 3 November 2020, and there were 74 respondents in total.[18] The majority of the respondents were between 30 and 59 years old and highly educated (see Table 12 and Table 13). It is clear that the respondents are not fully representative of the general population, and the responses may reflect the respondents' comparatively high level of education. It is possible that a survey on this topic, particularly one that was projected to take a relatively long time (15 minutes) to complete, was more motivating to more highly educated respondents who are familiar with academic research. In addition, the fact that project members shared the questionnaire link in their own networks made it quite likely that the link reached populations with high levels of education. While this means that the participants are not fully representative, the data still offers a useful glimpse into the opinions of potential subtitle users.

Age	Percent of respondents
18-29	16%
30-44	38%
45-59	32%
Over 59	14%

Table 12: Age of questionnaire respondents.

Highest level of education completed	Percent of respondents
Secondary education	4%
Vocational qualification	3%
Some studies at a university or other institute of higher education, but no degree	15%
Undergraduate degree	22%
Postgraduate degree	34%
Doctoral degree or higher	23%

Table 13: Education level of questionnaire respondents.

The respondents' Finnish skills and contact with Finland were quite varied. In the question about languages, most respondents reported not understanding any spoken Finnish at all, but the rest of them had varying levels of Finnish comprehension skills

[18] There were 74 responses to all multiple choice and likert scale questions, except for the first comprehension question after each video clip (*"Did you understand what the topic of the video was and what was said about that topic?"*), where one respondent had neglected to respond to any. The missing responses to the two questions were from different respondents.

(see Table 14). This distribution is quite beneficial for the questionnaire, as a significant majority of the respondents would not have been able to follow the video clips without subtitles. Similarly, a majority of the respondents, 68%, had never lived in Finland. The remaining respondents were living or had at some point lived in Finland (see Table 15).

Finnish comprehension	**Percent of respondents**
Not at all	58%
Some words and phrases in simple Finnish (CEFR A1)	15%
Main points in simple Finnish (CEFR A2)	10%
Majority of simple Finnish (CEFR B)	4%
Even complex Finnish (CEFR C1)	7%
No difficulty understanding any Finnish (CEFR C2)	7%

Table 14: Questionnaire respondents' comprehension of Finnish.

Living in Finland	**Percent of respondents**
Never	68%
Currently live in Finland and have lived there for 1-5 years	5%
Currently live in Finland and have lived there for 6-10 years	5%
Currently live in Finland and have lived there for more than 10 years	18%
Have previously lived in Finland	4%

Table 15: Length of time lived in Finland by questionnaire respondents.

All respondents reported watching some subtitled programming. The most frequent response, by 42%, was watching subtitled programming occasionally, while 38% reported watching subtitled programming often and 20% reported watching subtitled programming rarely. This selection of responses suggests that the respondents are familiar enough with the format to assess the subtitles from their subjective perspective, but a majority of them are not frequent watchers of subtitles, which may make it challenging for them to follow subtitles and could result in more negative responses.

8.3.1.3 User data collection

Focus groups

In all four focus groups, a small group of participants viewed the evaluation material and answered the moderator's questions concerning their comprehension and appreciation of the material, cognitive load caused by viewing the material, and general views on using automatic subtitles (see Appendix C for a full script of the focus group sessions; the

Finnish-language and English-language focus groups followed the same script, so only the English scripts have been included). In addition, the participants were encouraged to engage in dialogue with each other. They were also given the opportunity to complement their responses in a short online questionnaire after the focus group session. The questionnaire consisted of only two questions, one offering a chance to add comments on the topic of the focus group, and another to give feedback on the study (see Appendix C for the questionnaire; the questions in the English and Finnish versions of the questionnaires were identical, so only the English versions have been included).

While the groups largely limited themselves to answering questions, the English groups in particular initiated some interactions that offered additional insights into the participants' views. The data yielded by the focus groups was highly qualitative in nature, as the questions were open-ended and participants were allowed to answer in their own words and even discuss points beyond the topic of the questions. The discussions were also somewhat different from each other. They did not follow the prepared script exactly, but they all covered the four themes that were determined to be the focus of the study.

In the first round, the two focus groups were shown one approximately five-minute clip of a documentary program with fully automated MT subtitles, while the second round involved two shorter clips (see Section 8.3.1.1 for details). Because the participants in the second round were shown two clips, the questions in these focus group sessions were less detailed than in the first sessions, but they covered the same topics of comprehension, appreciation, cognitive load and attitudes towards MT. Keeping the questions more general was also decided to be a useful approach, because the first sessions showed that participants do not tend to give exact answers to very specific questions. The focus of the discussions was thus on general experiences and preferences, allowing for more informal commenting.

The quality of the subtitles was problematic and affected the nature of the focus group discussions. It was immediately obvious to the participants that the subtitles were not made by professional subtitlers. There were noticeable mistranslations, as well as issues with the language and synchronization of the subtitles. Therefore, the decision was made to signal beforehand that these are automated subtitles to avoid unpleasant surprises that might distract the conversation. In the first focus group discussions, the moderator was scripted to make the following comment in the introduction: "*[W]e are conducting research on the use of technological tools in translating and subtitling TV programs, making the process faster and more automatic. We are now trying to find out what viewers think of subtitles that have been created in a new way.*" However, even with this subtle suggestion that automation was involved, the first focus groups ended up speculating on the origins of the subtitles and were noticeably thrown off by the low quality. We therefore decided to be more explicit in the introduction of the questionnaire and in the second focus groups, by stating that the subtitles are fully automatic. While this decision comes with the risk of priming participants to respond according to predisposed attitudes towards machine translation, this was seen as the smaller risk, as it was an organic way to introduce the subtitles, and it allowed respondents to start with more realistic expectations. The difference in the introductions may explain why both of the first focus groups discussed the need to label machine-translated subtitles as such, while the second focus groups did not mention this issue, as the sample clips had already been labelled for them.

The focus group sessions were recorded on Google Meet's screen recording function, and they were anonymized and transcribed. All focus groups started with a warm-up question on how much subtitled programming the participants are used to watching. The purpose of the question was to act as an ice-breaker and to provide some background information on the participants' experience with subtitles. Then, the participants watched the subtitled video clip and after the clip, the discussion concerning the subtitles started. In the second focus groups, participants first viewed the first video (fully automatic news clip) and discussed it briefly, then viewed the second video (documentary clip with light pre-editing of ASR for MT) and covered the same topics as with the first video, and finally they discussed general issues related to MT subtitles and their viewing experience. At the end of all four sessions, the participants were given a link to the online questionnaire that they could submit if they wanted to add anything to their comments during the focus group, or if they wanted to give feedback on the session. One participant from each of the first two focus groups submitted comments via the questionnaire, two participants completed the questionnaire from the second English focus group, and there were no responses from the second Finnish focus group. All comments were positive about the study but did not add much of substance to the discussions.

Questionnaire

The purpose of the questionnaire was to follow up on the topics addressed in the first focus groups and collect quantitative data to complement the qualitative focus group data. Some topics turned out to be challenging to explore in the focus groups, so they received particular attention in the design of the questionnaire. In particular, comprehension of the subtitled clips and the cognitive load caused by viewing them were topics that focus group participants discussed quite vaguely, so the questionnaire contained several questions on these topics. In addition, there were questions about general appreciation of the clip, as well as questions concerning the respondents' thoughts on machine translated subtitles and potential uses for them (see Appendix C for the full questionnaire). Most questions were closed questions, either five-point likert scale questions or multiple choice questions. In addition, there were four open questions at the end of the questionnaire.

Our original intention was to launch two questionnaires, one in English and one in Finnish. For that purpose, two questionnaires were designed and video clips were prepared for both. However, as it became evident that the quality of the Swedish ASR combined with MT into Finnish was considerably lower than the quality of the Finnish ASR combined with MT into English, the decision was made to focus only on the English questionnaire. This decision was made after a small pilot study with students at the University of Helsinki, whose responses confirmed that reactions to the Finnish questionnaire would be rather negative. The decision was not taken to avoid collecting unflattering data, but to avoid conducting an evaluation that would have little use. With poor subtitle quality, it would have been difficult to elicit substantive answers beyond an overall negative assessment. For example, one respondent in the Finnish pilot study mentioned having “missed one video completely” because of problems with the subtitles. Such a viewing experience would not have provided meaningful data and could have resulted in disappointment and frustration for the respondents. It would also have been difficult to compare the data from the two questionnaires with each other due to the noticeable quality difference.

The English questionnaire was also determined to be more useful than the Finnish one because there is more need for Finnish into English MT subtitles at YLE. English-speaking audiences are not well served by YLE’s current provision of subtitling, so it makes sense to test it in more detail to explore the option of using automation to fill a gap in the language provision. Thus, testing the more mature system that answers an obvious, existing need was deemed to be the best use of time and resources at this stage of development. However, we decided to conduct the focus group sessions in the Swedish into Finnish language pair as well, because focus groups are a good way to collect preliminary information that can help with further development. The focus group setting provides a more useful context for presenting something unfinished, as it allows the researcher to explain and contextualize the material, answer questions and gauge participants’ reactions face to face. The nature of the focus group data, including its exploratory focus and its highlighting of attitudes, experiences and social dynamics, facilitates the early examination of systems that have not quite taken shape yet, whereas a questionnaire requires self-explanatory material that the respondent can react to without additional guidance. In addition, the second focus group offered an opportunity to compare fully automatic subtitles and subtitles with light pre-editing of ASR for MT, which was particularly relevant for the Swedish into Finnish language pair where differences between the two were greater than in the Finnish into English language pair.

The published English questionnaire contained two video clips (news and documentary) with fully automatic subtitles (see Section 8.4.1.1 for details). There was a series of identical questions on each video related to comprehension, appreciation and cognitive load. After the questions specific to each video, there were general questions regarding the respondents’ opinions on MT subtitles, and then the four open questions at the end of the questionnaire. It was decided that it would be useful to ask questions on two videos rather than one to gather more data, and to see whether some reactions are specific to an individual video.

The first of the four open questions asked respondents to suggest possible use contexts for automated subtitles. The second question asked for suggestions on what should be improved in the subtitles, and the third one solicited any other feedback on the subtitles. The final question asked for feedback on the questionnaire. The open questions were voluntary, but the respondents used them quite actively. There were 53 responses to the first question, 50 to the second, 29 to the third and 20 to the final open question.

8.3.2 Analysis of viewer data

In this subsection, we will discuss the findings from the three stages of the viewer study in chronological order. We will begin with the first round of focus groups, then we will discuss the questionnaire data, and finally the second round of focus groups.

8.3.2.1 The first focus group discussions

The focus group discussions covered all four themes of the study, and they provided substantive information on participants' views. The immediate reaction to the subtitles was negative in both groups. In the Finnish group, the first comment was that the subtitles were "sub-par", contained errors and were difficult to follow. Subsequent comments mentioned problems with segmentation as a noticeable issue. In the English group, the discussion started more cautiously with a negative comment about the visual appearance of the subtitles, but quickly progressed to complaints about the subtitles being "ropey", confusing and difficult to follow. However, the immediate assessment was not exclusively negative in either group. In the Finnish group, one of the first commenters mentioned not noticing any of the errors brought up by another participant and thinking that even the most significant translation errors were intentional humour. In the English group, one participant's first comment was "I was really happy to see Finnish[19] subtitles because possibly I would never get to see such content otherwise". Still, the discussion about the quality of the subtitles was predominantly negative in both groups, with multiple criticisms over obvious mistranslations; clumsy, unclear and unidiomatic or even nonsensical language; fast or uneven pace of the subtitles; lack of synchrony between the subtitles and the spoken language; and poor segmentation and line breaks.

While the quality issues clearly affected the participants' viewing experience, there were comments in the English group that expressed a sense of acceptance towards automated subtitles as better than nothing and an improvement over the current situation, where it is difficult to find English subtitles for Finnish programs. Participants in the English group also commented that it is possible to become used to this type of subtitling and to learn to work out the meaning to the extent that the subtitles would be helpful in everyday situations. In other words, both groups noticed similar quality issues and considered them quite serious, but the English group was more prepared to overlook those issues to fulfil an urgent need for translated content. One participant in the English group clarified this need for linguistic access in personal terms, stating that this kind of subtitling "gives you a lot more access, a lot more, ability [...] to actually take part in the cultural events. There are these things you'd like to talk about with Finnish colleagues and friends the next day," This difference between Finnish-speakers and English-speakers was a significant dynamic in the audience evaluations: whereas Finnish participants are well served by Finnish-language content and professional translations, English-speakers have little access to Finnish news and current affairs programming or any other Finnish-language content, and it is therefore easier for them to imagine using even low-quality subtitles. The English-speakers' willingness to tolerate imperfect subtitles may have been supported by the fact that the quality of the English subtitles was better than that of the Finnish subtitles.

[19] The participant mistakenly refers to 'Finnish subtitles' here, even though the subtitles are in English throughout the video. However, the participant's clear intention is to discuss subtitles translated from Finnish into English. The same participant had mentioned earlier in the discussion that they had tried to find Finnish content translated into English, but that had been difficult, and this comment is referring back to that remark.

Despite quality issues, both clips were fairly understandable to the participants. Although they reported some confusion over the subtitles, they appear to have been able to understand the main gist of the clip. Several participants in the English group stated that they were able to follow the narrative, and some described the general sense of the video fairly accurately. They did, however, express concern about the subject matter coming across as heavy and difficult, which may have been, in part, due to faulty subtitles. Similarly, in the Finnish group, the participants described some confusion and difficulty following all the details of the narrative, but several participants stated that they were able to understand the clip at least partially, and one participant described the contents of the clip quite accurately and to some detail.

The sense of confusion, which was frequently expressed in both groups, suggests that the readability of the subtitles was not optimal, and that reading the subtitles while trying to watch the clip caused a heavy cognitive load. Comments in both groups suggest that the viewer is forced to pay closer attention to the subtitles on one hand to understand what they say, and to the rest of the program on the other hand to fill gaps left by the subtitles. The viewer has to navigate between the various meaning-making modes of the program, use one's imagination to interpret the meaning of the narrative, check facts from other sources, and perform a variety of other mental tasks. This kind of viewing is more taxing than a regular viewing experience. Poor readability also makes subtitles more distracting than in a regular viewing situation. As one participant in the English group put it: "As you concentrate on the subtitles, you start missing something on the video itself and vice versa, if you're looking at the video, you might miss something important in the subtitles." Another commenter in the English group described the viewing experience in the following way: "I don't think this is that kind of a program you could sit and just half-watch with one eye, while you are scrolling on your phone. So you have to be totally focused in order to keep the thread." For viewers used to multitasking, it can be problematic if subtitles force them to focus all of their attention on trying to understand the program. A related concern may be that these subtitles may limit viewers' enjoyment of the program. As one participant stated in the Finnish group: "This was not meant to be enjoyed, just kind of run through somehow."

Because the viewers' experience of the subtitles was not particularly positive, ideas for potential uses for the subtitles were limited. If the viewing experience is not enjoyable, it would make sense to limit the provision to content that is not watched for enjoyment but for a functional purpose, such as gaining important information. This was how many focus group participants responded when asked about potential uses for automated subtitles. As the general attitude in the Finnish group was more negative, there was also less enthusiasm for suggesting potential uses. However, as the conversation progressed, some participants softened their views. One explicitly declared revising their original negative comment and said that "if it was some kind of like announcement or some sort of, like, news story or something like that, some force majeure situation that needs to be shared with a lot of people immediately, I could watch that, but then with factual programs that I would watch seriously, those subtitles would bother me there." This was a broadly shared view in the Finnish group, although one participant also remarked that factual errors in the subtitles may cause problems if the subtitles are used to convey important and urgent information.

The participant quoted above as becoming more accepting of automated subtitles also suggested another possible use. This suggestion concerned topics of personal interest, such as hobbies which may originate in a country and language not familiar to the person, such as Japanese. If the topic is not broadly popular in the target country, there may not be professional translations, and automated translations could provide access to those who are interested in the topic. In this case, personal interest provides a strong enough motivation to follow subtitles even when the quality is low.

The English group spent more time discussing potential uses for automated subtitles, demonstrating that the issue is more urgent for them than for the Finnish participants. Again, breaking news and current affairs were mentioned as a potential use for automated subtitles, as well as local news and information that was difficult to access in English. The access provided by automated subtitles was seen as especially important in the situation these focus group participants were in, as they did not understand the languages of the country in which they live well enough to follow the local media. Therefore, they had an immediate personal motivation for using even imperfect subtitles. They also mentioned other possible uses, including lighter, entertaining content where exact accuracy is not very important. As one participant put it, “As long as I am not signing bank documents, I don’t really care if it’s not a 100 percent accurate translation.” One participant even suggested that automated subtitles could be introduced to all programming. That would give the audience the choice to either use them or not, and using the subtitles could both encourage increased media use and benefit audiences as a tool for learning the language. However, some doubts were expressed about the reliability of automated subtitles as a source of information in high-stakes situations, such as COVID-19 guidelines, or as a reference point in language learning. It was also pointed out that for important cultural products, such as classic Finnish films, automated subtitles would not be suitable. Nevertheless, the repeated sentiment was that this would be better than nothing and would facilitate access and participation in society.

To explore the groups’ willingness to use automated subtitles further, they were asked whether they would prefer immediate access to automated subtitles or delayed access to professional subtitles created by humans. The views were mixed and most answers were not straightforward. The exchanges reflected an appreciation for both human skill and machine efficiency. In the Finnish group, only one participant stated a straightforward preference for immediate machine-translated subtitles, and another stated a straightforward preference for human translation. Others expressed various caveats, such as preferring post-edited machine translation, or requiring a clear label stating that the subtitles are automated. In addition, the preference is often context-specific, and while the participants may be willing to view urgent news with automated subtitles, they usually have a preference for human translation in other contexts.

In the English group, the same ambivalence was evident, and most respondents stated that their preference depends on the context. Only one expressed a straightforward preference for human translation, while one other first preferred machine translation but later agreed with the rest of the group and stated that while machine translation is useful in some contexts, there are times when accurate and reliable human translation is needed. Human translation was put forward as the gold standard, and the quality of the automated subtitles raised concerns, but the immediate access it provides was rated as an important factor. The conversations demonstrate that automation and professional human quality both have their place in consumers' preferences, and the immediacy of automated translation makes it an acceptable, even if not ideal, solution for some contexts. However, as the focus group participants were quite vague in their comments, it was decided that the questionnaire should probe this topic further with a multiple-choice question.

8.3.2.2 Survey

The focus groups provided substantive information about the participants' relationship to subtitles, but some of the data can be vague, as participants use their own words rather than putting responses into exact categories. Therefore, it is helpful to complement open-ended focus group data with the questionnaire's more neatly categorized responses. The questionnaire responses largely follow the patterns set by the first focus groups. Respondents indicated a negative view towards the subtitles but understood much of the content and did not reject the possibility of using automated subtitles in some contexts. Figure 17 gives an overview of viewer responses to the 6 Likert scale questions that were asked of the two videos separately. A score of 5 is always the positive end of the scale, while a score of 1 is the negative end. The full questions and response scales are:

- Did you understand what the topic of the video was and what was said about that topic?
  1 not at all – 5 yes, completely
- Was the video pleasant to watch?
  1 not at all pleasant – 5 very pleasant
- How useful were the subtitles in helping you understand the clip?
  1 not at all useful – 5 very useful
- Did the information in the subtitles seem accurate to you?
  1 no, not at all accurate – 5 yes, completely accurate
- In comparison to an average experience of viewing a subtitled program, how well did you manage to read the subtitles all the way through?
  1 considerably worse than usual – 5 considerably better than usual
- In comparison to an average experience of viewing a subtitled program, how much mental effort did you have to invest to learn new information from the clip?
  1 considerably more than usual – 5 considerably less than usual

In the figure, 'video 1' refers to the documentary clip, and 'video 2' to the news clip. In addition, the questionnaire contained three multiple choice questions on the contents of each video, and general questions about respondents' views on MT subtitles. Those questions will be discussed below.

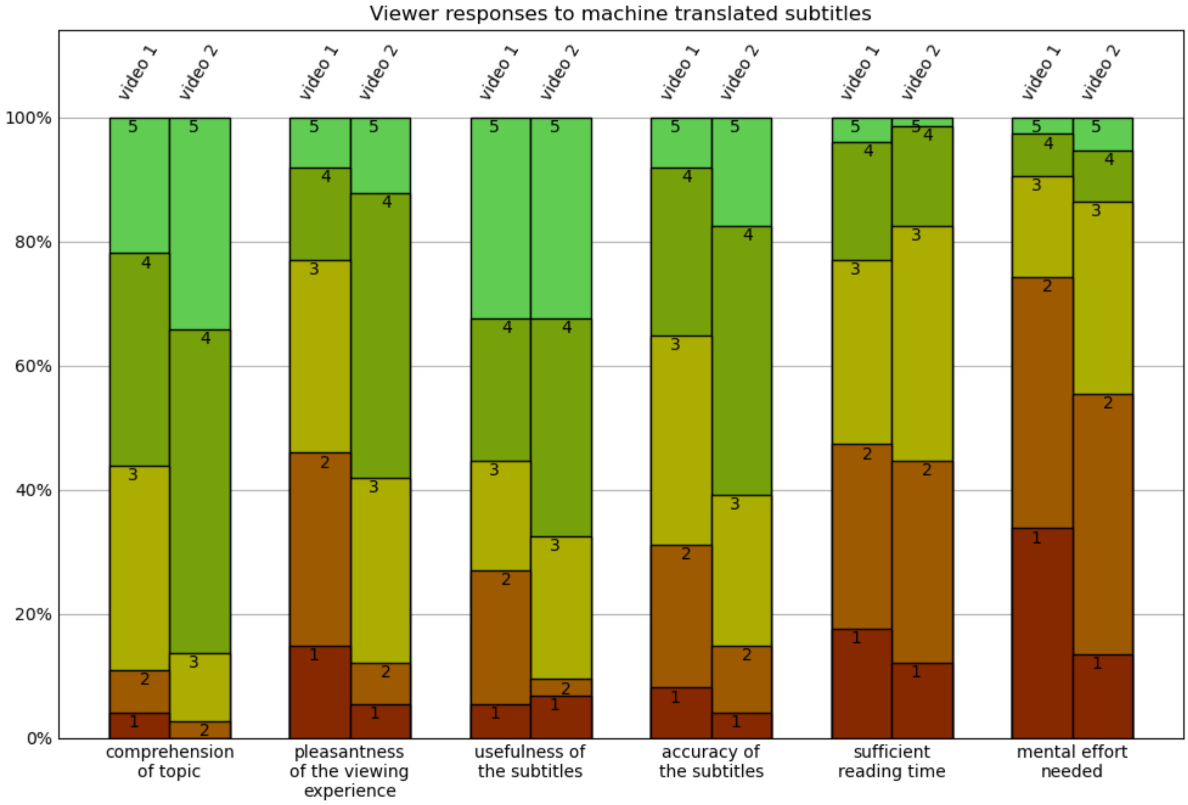


Figure 17: Viewer responses to Likert scale questions on the two MT video clips.

As can be seen in Figure 17, the responses to questions on the viewing experience and comprehension suggest that the quality of the subtitles was somewhat acceptable, even if the response was not overwhelmingly positive. When asked whether the video was pleasant to watch (the second question in Figure 17), the average score for the documentary clip was 2.7, and for the news clip the average was 3.5. While the experience does not appear to have been an exclusively enjoyable one, the majority of the responses were 3 or higher, and the news clip stands out as the more positive experience. When asked how useful the subtitles were, the answers were even more positive. On the documentary clip, the average score was 3.6, and the news clip again received an even more positive score of 3.8. This may be indicative of the fact that most respondents did not understand any Finnish and needed the subtitles to understand anything, but it does also suggest a successful experience where the respondent was able to rely on the subtitles for information. In the question on the perceived accuracy of the subtitles, responses to both videos were again in neutral or positive territory: the average score for the documentary clip was 3.0 and for the news clip 3.6.

On the questions related to cognitive load, responses were not as positive as on the viewing experience. The first of these questions tested reading times by asking how well respondents managed to read the subtitles all the way through in comparison to a normal viewing experience. The average score was 2.6 for both clips. It appears that for a reasonable proportion of the respondents, the reading speed felt similar to a normal subtitle reading situation, as a 3 was the most popular choice in both cases, chosen by 30% and 38% of respondents respectively. For those who rated the experience negatively, it was often only moderately more negative than a normal situation, with 30% and 32% of responses respectively at 2, and only 18% and 12% of the responses respectively at 1.

The second cognitive load question asked about the mental effort needed to learn new information from each clip. These responses indicate clearly that these video clips were more mentally taxing than professionally subtitled clips. The average score for the documentary clip was 2.0, and as many as 41% of the respondents chose 2, and 34% chose 1. For the news clip, the results were again better but still quite negative: the average score was 2.5, and 42% of respondents chose 2 and 31% chose 1. These numbers demonstrate that even if the quality issues in automated subtitles do not severely hinder understanding, they make the subtitles and therefore the entire program difficult to follow.

The respondents' sense of cognitive load was also tested with a question posed jointly for both clips (see Figure 18), which asked more broadly whether this viewing experience felt more difficult than normal experiences with subtitled content. The 1 to 5 scale on this question ranged from "considerably more difficult" to "no more difficult". The responses were again towards the negative end of the scale, with an average score of 2.4. These responses again indicate that there is a severe cognitive load associated with automated subtitles. It should be noted, however, that the response scale on this question was phrased differently than the other cognitive load questions. Whereas the scales for the other cognitive load questions ranged from better than usual to worse than usual, here respondents did not have the option of responding that this experience was less difficult than a normal viewing experience. This was done to gauge the extent of the negative reactions even further, because it was realistic to expect that a majority of the cognitive load responses would be negative due to the perceptibly low quality of the subtitles. Indeed, the negative reaction turned out to be quite severe. There is, however, a risk that some respondents misread the options and assumed that the scale was identical with the previous questions, which may explain some of the responses, but the responses should still be taken as an indication that cognitive load is a serious issue.

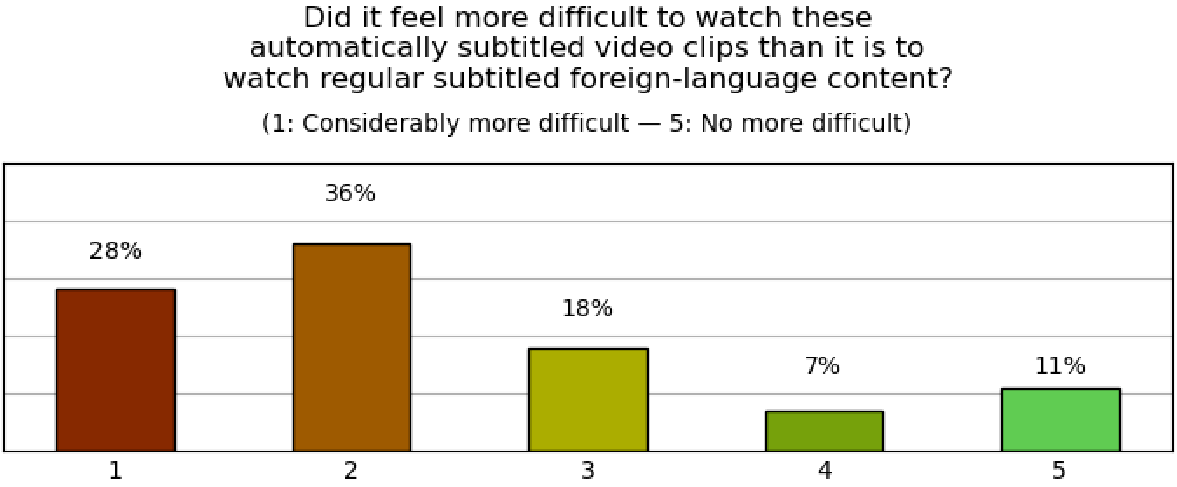


Figure 18: Viewer responses to Likert-scale question about cognitive load.

Despite this long list of quality complaints, questions which probed respondents' comprehension indicate that the video clips were fairly understandable. In the first question displayed in Figure 17, the average score for the documentary was 3.6 and for the news clip 4.2, so the response to the news clip was again better, but both scores are quite positive. There may, however, be some difference in comprehension between those who have some contact with or knowledge of Finnish and those who do not. All respondents who chose 1 or 2 in these comprehension questions had never lived in Finland and reported not knowing any Finnish at all. However, the 1 and 2 responses were clearly in the minority even in this subgroup, so the difference is not significant. This finding does suggest that requirements for MT quality should be even higher when the source language and culture are less familiar to viewers, and even limited knowledge of the source language may play a role in comprehension.

The same tendency towards comprehension was evident in multiple choice questions on the contents of the clips. On the documentary clip, a vast majority of responses were correct in two of the tree questions. The third question was more difficult and was related to a segment in the clip that was somewhat confusing both in the subtitles and in the narrative itself, and on that question, 53% of respondents chose the option "I don't know", while only 11% of responses were correct. In the news clip, either a plurality or an outright majority of responses were correct on all questions. These findings confirm that viewers both feel like they have mostly understood what they saw and also substantively answer comprehension questions fairly well.

While the above analysis reflects some positive signals about the quality and acceptability of MT subtitles, the open questions at the end of the questionnaire tell a slightly different story. The second open question asked respondents to name some things they feel should be improved in the subtitles, and the large number of responses (50) indicates that there are many quality issues that require attention. Many responses mentioned more than one quality issue. The largest category is the language of the subtitles (21 comments in all), including complaints such as interference, clumsiness and lack of fluency, and issues with grammar, syntax, spelling and punctuation. This category can be further divided into specific complaints about grammar, syntax, spelling and punctuation, which comprise approximately half of all the language issues (10 comments), and more general comments of clumsiness, interference and awkwardness (11 comments). The second largest category of quality issues is segmentation, which was mentioned 16 times, such as in the following comment: "The sentences were sometimes cut in weird places. Try improving it by showing the whole, or as much as possible of one sentence at the time, instead of cutting it in parts. That gives better flow of reading." The next largest category related to the quality of the translation, which received 11 comments. The comments tended to be quite general and simply mention that the translation was not accurate or reversed meaning in some places. Some respondents pointed out specific instances where the translation was not accurate in their estimation. After translation quality, the next largest were two categories that contained ten responses: visual aspects of the subtitles and pacing and reading speed. While this study was not intended to explore the details of visual presentation, it is useful to note that the readability of subtitles is significantly affected by visual factors such as the size and placement of the text and the color of the background. In addition, several commenters complained that the subtitles cover crucial information, which highlights the importance of taking the visual context into consideration. The issue of pacing also affects readability, and commenters pointed out that they did not have enough time to read all subtitles fully, either because they felt that the subtitles were not on screen for long enough, or because the text was difficult to read and would have required more time. In the next category, there were nine comments in total. This category consisted of comments that wanted to see more human involvement in the subtitling process, either in the form of post-editing or creating the subtitles from scratch.

In addition to these large categories, there were a few comments that pointed to problems with timecoding and synchronization, condensation or omission, excessive cognitive effort, general issues with the technology, and simply "many things" in need of improvement. One comment suggested that the subtitles lacked the kind of "feeling" that is characteristic of good subtitles. In all, the open answers reveal a wide range of challenging areas which affect readability, comprehension and trust, and many echo topics that were mentioned in the focus groups. In addition to the responses to this open question, some of the comments in the third open question, which invited any other feedback on the subtitles, returned to the topic of quality. Those responses included four remarks about the quality of the language, three about typography and visual aspects of the subtitles, two about translation quality, two about pacing and reading speed, and one about condensation and omission in the subtitles. Taken together, all these responses show that there are serious shortcomings in the quality of these automated subtitles, and viewers do pay attention to them. However, it should also be pointed out that 24 respondents did not give an answer to the question about quality issues, which may suggest that they were reasonably satisfied or at least could not think of any issues worth mentioning.

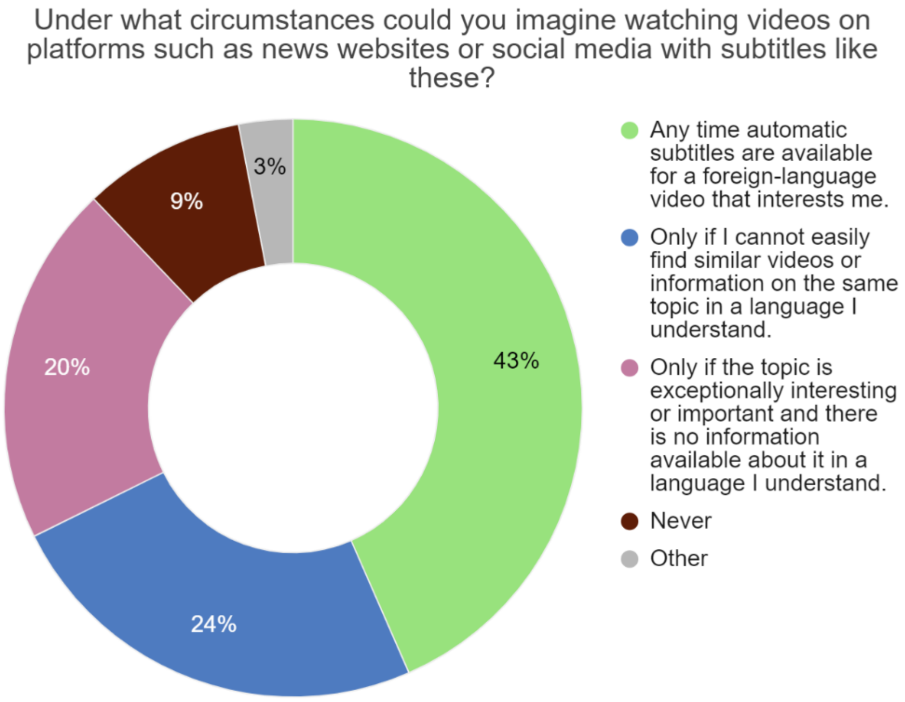


Figure 19: Circumstances for watching videos with automated subtitles.

The questionnaire also asked respondents to consider possible contexts where automated subtitles may be useful. This was examined through a multiple-choice question and an open question. The multiple-choice question asked respondents to choose under which circumstances they might watch automatically subtitled videos, with the options becoming gradually more restrictive (see Figure 19). A plurality of respondents, 43%, chose the least restrictive option, and the percentage of responses decreases towards the more restrictive options. Respondents were also given the opportunity to answer “other” and explain their view, and two respondents chose that option. The first response was: “Only if I already partly understand the spoken language (but not quite well enough to use its non-translated subtitles), and it is an important topic with no other source of information”. The second response was: “For educational/research purposes; if the quality of auto-subs is low (as has been reported in the case of Greek subs on Amazon Prime, for example), I wouldn't consider relying on them for information or entertainment purposes”. These responses suggest that there is motivation for using automated subtitles for some purposes, and the open question further illuminated what those purposes might be.

The open question gained a solid 53 responses. Nine of them simply answered “no”, that they cannot or probably cannot think of uses for automated subtitles. One answered ‘yes’ without explanation. Other responses contained suggestions for genres and contexts where automated subtitles could be used. The most frequently mentioned genre was news (e.g. local news, breaking news) and current events, which was mentioned 19 times. Admittedly, that may be due in part to the fact that the video clips shown in the questionnaire were in this genre. Other genres only received individual mentions, and they include government announcements, emergency reports, weather, traffic, sports, anime, k-pop, films, tv shows, documentaries, podcast videos, and interviews and discussions. What stands out is that many of these genres relay time-sensitive information, and as such, the immediacy of automated subtitles is a natural fit for them. The focus on live or urgent information was indeed present even in many comments that did not mention any genres specifically. In total, six responses mentioned live content, and 19 mentioned content that was not necessarily live but was urgent or time-sensitive, where automated subtitles can make important information available quickly.

On the other hand, a few comments presented an opposing view: four respondents pointed out that automated subtitles might work for low-stakes content where accuracy is not crucial, and one suggested it might be useful for old content that would not be important enough to get translated otherwise. This comment is also an example of a process or resource-related rationale for using automated subtitling. In addition to this comment, six others mentioned that automated subtitles could be a backup solution in a situation where there are no resources for human translation, and two respondents suggested that they could be used as a first step for post-editing, or while waiting for a human translation. It was also suggested that automated subtitles could be used if the material is simple enough to be manageable by the tools. In addition, automated subtitles were described by two respondents as a support element, a “sanity check”, for viewers who understand some of the source language but not all of it. Finally, four respondents suggested that automated subtitles could be used for YouTube videos that cannot be translated by humans. As these examples demonstrate, there is a wide range of possible uses for automated subtitles, but they appear to be primarily preferred as a support system when nothing else is available and circumstances require fast delivery of information and content.

As in the focus groups, the questionnaire respondents were asked to decide whether they would prefer immediate machine translation or a human translation with a delay. This was a multiple-choice question (see Figure 20). The responses were spread among all options, but the human-translation options were slightly favored: 26% would definitely choose human-made subtitles, and 30% would choose human-made subtitles in most cases. There were also three "other" responses, which all stated in various ways that the choice depends on the context.

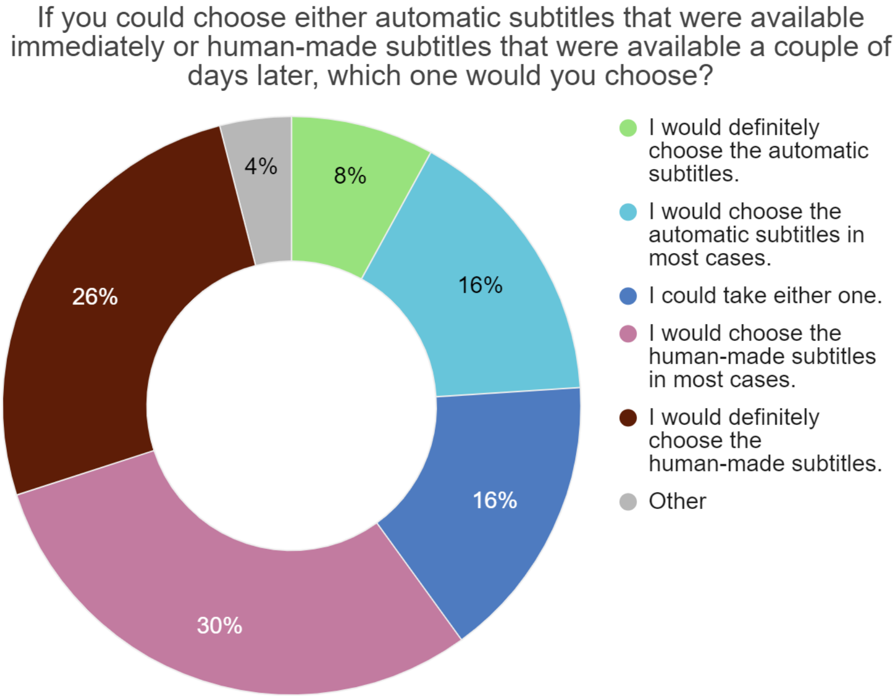


Figure 20: Choice between machine-translated or human-made subtitles.

There were also several responses in the open questions which expressed a preference for human involvement in the subtitling process. As was discussed above, human involvement was mentioned in the second open question as a way to improve subtitle quality. In addition, there were nine responses in the third open question, which invited other comments on the subtitles, that suggested that human involvement would be preferable to fully automatic subtitles. Furthermore, four responses to the third question stated that the technology needs improvement, which could be taken as an implicit statement of preference for humans over automation. In conclusion, there seems to be willingness to use automated subtitles, but in specific and limited circumstances, not to replace human translation but to intervene where it is not available. As in the English focus group, nine commenters in the third open question stated that subtitles like these would be better than nothing: helpful, useful and good enough. In addition, a few comments in the question about feedback on the questionnaire expressed their thanks for the initiative. There is clearly a need for and an interest in automatic translation, but on the whole, human translation appears to be preferred if the option exists, even if it means that the content is slightly delayed.

8.3.2.3 The second focus group discussions

The themes of the second focus groups were the same as in the first focus groups and the questionnaire. The only major difference was that both groups were shown two video clips, one of which involved light pre-editing of the ASR output prior to the MT phase. The general sense of the discussions was similar to the first focus groups and the questionnaire. Participants in both groups expressed skepticism over the quality of the subtitles. A participant in the Finnish group equated it to Google Translate and even added that it resembled "bad Google Translate", while a participant in the English group said the subtitles resembled the broken English of the fictional character Borat, and another described them as Google Translate "with like five different languages". Participants in both groups criticized obvious and amusing mistranslations as well as awkward language and grammar issues which were detrimental to readability. A participant in the Finnish group also called the subtitles monotonous and devoid of nuance. Problems with fast pace and segmentation were also mentioned in both groups. In the Finnish group, the pre-edited clip received a more positive assessment than the fully automatic clip. It was described as more fluent and understandable, better in synchrony with the video, and having a calmer pace. However, some participants commented that the subtitles were still somewhat confusing and contained strange words. Thus, even though the participants were able to notice a difference between the two clips, the improvement was not sufficient to make the pre-edited subtitles easily acceptable.

Similarly, participants in the English group found the pre-edited subtitles slightly better, even if still problematic. They mentioned, in particular, that the pace was better than in the first video, and there was more time to read the subtitles. However, this could be due to the fact that the pace of the documentary clip was slower to begin with. When the pre-editing was mentioned, some participants claimed they noticed a difference between the two clips and ascribed that to the manual editing. However, the differences between the two English translations were not very noticeable, so it cannot be assumed that the participants' perceptions were predominantly due to pre-editing. Nevertheless, the idea of manual editing was well received. Participants in the English group also mentioned visual aspects of the subtitles, such as font size, as concerns, and expressed a wish that on-screen text should be reproduced in the subtitles. Finally, as with other English respondents, this focus group was more accepting towards automated subtitles than the Finnish group, with repeated comments about how it would be "better than nothing", "definitely a help", and that even the quality is "getting there".

The discussions on comprehension followed a very similar track as in the previous focus groups and the questionnaire. Participants were able to describe the gist of the clips, although there were numerous comments about the clips being difficult to follow and requiring active interpretation and guesswork. The participants' descriptions of the contents of the clips contained some errors in both groups, but the main elements of the clips seem to have been understood. In the Finnish group, the slightly pre-edited clip received considerably more positive comments on comprehensibility than the fully automatic clip. In the English group, on the other hand, the difference between the two clips in terms of comprehension was less noticeable. In fact, several commenters mentioned having trouble with the pre-edited documentary clip because of the convoluted narrative and only seeing a fraction of it. One problem with the documentary was that the name Patria may not be familiar to English-speakers, so it is more difficult to follow the narrative than for a native speaker who immediately recognises the name. One participant even mentioned googling Patria during the session to understand the clip better. This is a fundamental challenge with machine translation, because it is not inclined to provide explanatory translations for concepts that are more familiar in the source culture than in the target culture.

Due to issues with quality and comprehensibility, the participants' comments suggested a rather heavy cognitive load, in a very similar way as in the first focus groups and the questionnaire. The effort required to view the clips was a topic to which the groups returned frequently. Both groups reported that the viewing felt heavy and demanding and required a significant amount of mental effort to fill the gaps left by the subtitles. They also suggested that it would be difficult to watch longer programs with subtitles like these. As one commenter put it in the Finnish group: “It was really exhausting. It felt like you really had to work hard to be able to follow it.” Similarly, a commenter in the English group said: “in the end, you still get the main picture of what's happening. It's just really, really hard. Like, you really put all of your energy, just to understand that main point.” Another commenter in the English group stated that “to watch any amount of time, of that, would be very hard on the head.” These comments make it clear that cognitive load is a significant issue, and it affects the viewing experience to a great extent. Another factor contributing to cognitive load is a sense of distraction caused by the subtitles. This was also reported in both groups. There appear to be at least two kinds of distraction: first, the subtitles require so much attention that they make it difficult to follow the program itself, and second, strange, awkward or unintentionally amusing errors in the translation act as a distraction by breaking the narrative flow and focusing the viewer's attention on the error rather than the substance of the program. One participant in the Finnish group even commented that “I would kind of like to follow the image as well”, which should, of course, be the default viewing experience.

The fact that these subtitles require more cognitive effort means that viewers have to be more motivated to use them than they need to be for professional subtitles. The focus group participants suggested that they would use them with a program they are very interested in or feel they need to understand. This was reflected in some responses concerning possible use contexts. Again, the Finnish group was reluctant to suggest possible uses at first, only mentioning that the program would have to be very interesting, or that they might use the subtitles just to see amusing errors or out of curiosity. However, as the discussion progressed, some specific suggestions were made, such as sudden, dramatic news stories like the 9/11 terror attacks or the Indian Ocean tsunami in 2004, when understanding the first news bulletins is particularly meaningful. The exchange on possible uses demonstrated one of the benefits of the focus group context, as participants visibly gained ideas from each other and started suggesting additional use contexts, such as emergency messages, or watching entertainment content in local languages when travelling. One participant told of an experience travelling abroad and seeing a local news story about the death of a celebrity but not understanding what was said, which presented a concrete case where automated subtitling would have been useful. Thus, the Finnish focus group was able to come up with contexts in which they considered automated subtitles genuinely useful, and news and current affairs were again the prominent genre.

The same was true of the English group. The participants were initially cautious about suggesting potential uses and expressed doubt about the usefulness of the system, but eventually they started proposing some ideas. The ideas were divided into two camps. First, some participants mentioned high-stakes, urgent and important news stories, such as the COVID-19 pandemic and its containment measures, an imagined declaration of war by the Finnish president, or other similar dramatic news. One participant described how uncomfortable it had been to not understand the government's COVID-19 information: "whenever the government had a, you know, a briefing on the latest situation, and, guidelines to follow and, you'd kind of have to wait for many hours, to get, some translation. And it was only a few sentences. So, you're kind of sitting here, thinking, okay, what do I do? Right now?" The current global situation thus provides an example scenario where information is needed urgently in multiple languages, and automated subtitles can help. On the other hand, the second type of suggested uses is the opposite in significance: light entertainment where accuracy is not as important, such as morning shows, celebrity interviews, talent shows, music programs or other "everyday silly things", as one participant put it. In addition, one participant mentioned how useful live subtitles would be in multilingual video conferences or similar events. Finally, just like in the first English group, one participant suggested offering the option of automated subtitles for everything. This list shows that although the group was cautious at first, they eventually presented a number of plausible scenarios and expressed a genuine willingness to use automated subtitles.

When presented with the choice of using either automated subtitles immediately or human-made subtitles a few days later, the answers in both groups were ambiguous but appreciative of human input. In the Finnish group, the balance was very much towards human translation, and no one expressed support for immediate machine translation. Two respondents did mention that they would use it in situations where they have to, but that they prefer human translation, and one commenter voiced support for post-editing the MT. In all, there was a cautious attitude towards automated subtitles. Towards the end of the discussion, one participant mentioned that in principle, it sounds like a good idea, but it still needs development and "it clearly still needs the human there to check the subtitles through." In the English group, the typical answer was to accept automated subtitles for certain genres or contexts, such as breaking news or light, unimportant "silly things" as discussed above. Five participants suggested that they would be receptive to automated subtitles under some circumstances, but they also mentioned that they preferred professional or post-edited subtitles in other situations. One participant expressed particular patience and motivation to wait for professional subtitles: "Although sometimes, if, let's say if it's a documentary, I could wait months for professionally made subtitles, over this." One participant, who did not explicitly favor either option, stated that the machine translation technology clearly is not "there" yet, suggesting a preference for human translation. One of the participants did not answer the question directly but tended to agree with others' views. The group was largely in agreement that machine translation has potential in some contexts, but human input still plays a valuable role.

To conclude, the second round of focus groups confirmed impressions from earlier stages of the evaluation. The discussions reinforced the idea that providing news and other important content, as well as some lighter materials, to English-speakers living in Finland would be a valuable service. However, as automated subtitles are still mentally taxing to use, they do not appear suitable for all contexts. While both language groups are somewhat critical of them, the English-speaking group is more open to using them when nothing else is offered.

8.3.2.4 Final Thoughts

All viewer evaluations provided a uniform picture of audience views on MT subtitles: the subtitles are fairly comprehensible, but their quality is problematic. While viewers express some level of acceptance for the subtitles, there is still work to do to reach genuinely usable levels of quality. Most crucially, poor quality causes high cognitive load and makes it laborious to follow the subtitled program. MT subtitles can be used if no other options exist, but most viewers prefer subtitles made or post-edited by professional translators. However, there was a noticeable difference between Finnish-speaking and English-speaking participants in how they rated the potential usefulness of the MT subtitles. Finnish-speakers saw fewer potential uses for automated subtitling, as they are already quite well served by Finnish media. English-speakers, on the other hand, saw it as better than nothing and a potentially helpful way of following Finnish media, because English subtitles are currently not often provided for Finnish programming.

9 Evaluation of Epic 6.10: Auto-generation and correction of content descriptions

The ongoing multimodal content revolution, which increases the need for audiovisual (AV) content descriptions to manage archival re-use and resale, was the main driver for the development of the AV content description component of the MeMAD platform prototype. As was highlighted through the work conducted in other MeMAD Work Packages (WP2, WP3, and WP5), current automatic methods of content description often require human intervention. This is particularly relevant for automatic video description. Whilst automatic methods of video captioning and visual storytelling were progressed beyond the SOTA in MeMAD, through multimodal integration and by combining computer vision modelling with models of human engagement with multimodal narrative, human input remains a crucial dimension in audiovisual content description. The development of MeMAD's content description prototype application was guided by earlier work analyzing wholly human-generated workstreams. As a result, this component provides an editing platform which draws together automatically generated video description captions, outputs of named-entity recognition and other automatic metadata generation processes, as well as transcriptions and translation services, for further processing by a human operator. A screenshot of the completed functional application prototype is depicted in Figure 21.

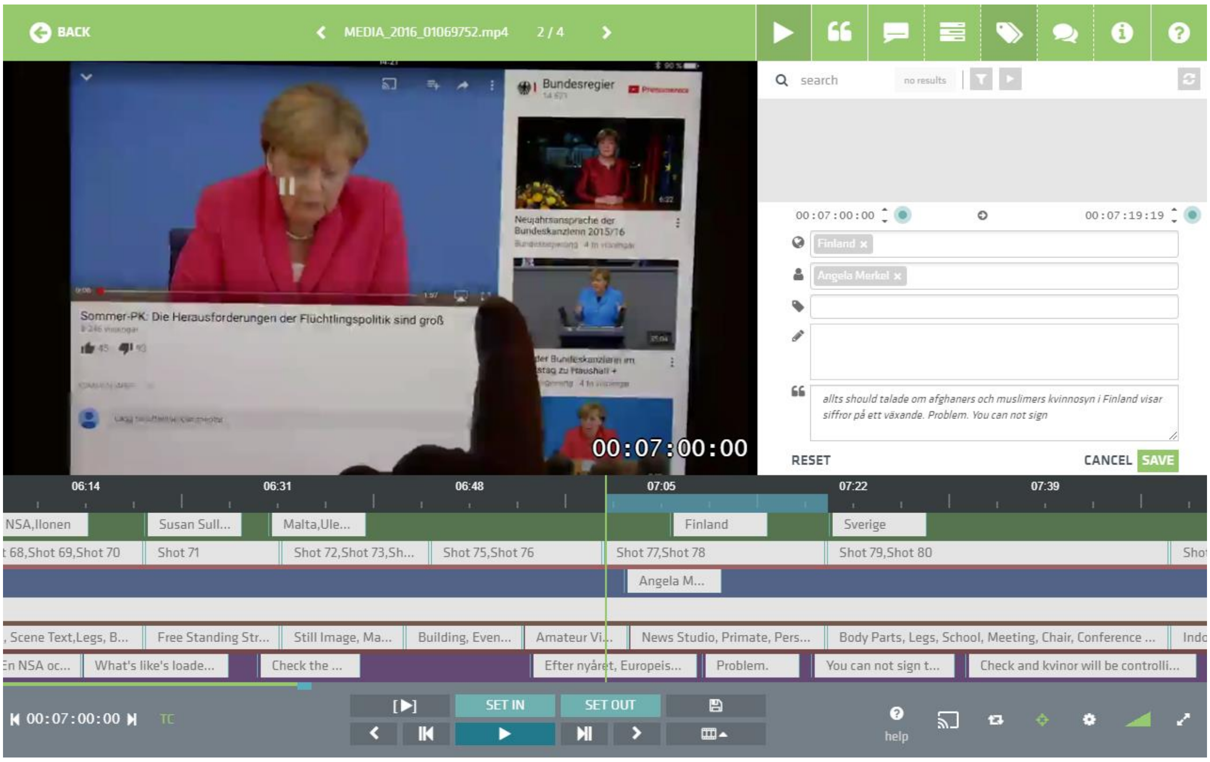
In line with this, the evaluation reported in this section focused on exploring and evaluating the prototype tool's affordances developed as part of Epic 6.10. As explained in D6.8 (Section 6.7), these affordances consist of the content description prototype application developed as part of T5.4 and its integration in the larger MeMAD media production platform along with the AME services from WP2 and WP3 which provide its input. In all, this functionality delivers the production of AV content descriptions based on a human-in-the-loop workflow that involves selecting, combining and post-editing automatically generated content and metadata.

The questions we sought to address in evaluating the developed functionality were:

1. In the perception of professional AV content describers, to what extent the prototype supports the creation of functional video descriptions?
2. Which of the tool's features are perceived to be most/least beneficial in this process?

Several aspects of the prototype, including usability, key features and the overall experience of working in this environment were explored through a mixture of methods including a user experience survey, focus groups, non-participant observation and multimodal analysis of observed work sessions.
As such, this evaluation serves as a natural extension to the design and development process started in developing the content description tool (cf. D5.4), in which first a requirements study was completed, feedback was gathered based on a first design, which in turn led to a second design that was again evaluated (on paper) by archival experts before being finalized as the blueprint for actual prototype development. This evaluation represents the next stage of assessment of the design, with a fully functional application

and a qualified user test panel with the intend of gauging how the application performs in real-life and whether crucial design decisions actually perform as intended.



00:07:00:00 — 00:07:19:19

Finland ×

Angela Merkel ×

allts should talade om afghaners och muslimers kvinnosyn i Finland visar siffror på ett växande. Problem. You can not sign

Figure 21: Screenshot of the evaluated prototype content description authoring application.

The evaluation consisted of using the prototype application to create AV content descriptions in such a way that they would be added to an archive system for later re-use. This involves describing content segments with keywords, named entities or even describing actions, people and places using full textual sentences. The workflow implemented by our prototype starts from a variety of metadata extracted from the content automatically and then lets users take over, reusing AME metadata where they see fit to deliver a complete content description.

We did not evaluate the quality of the machine descriptions or the users' output, nor did we measure or otherwise evaluate the speed of production. This was for several reasons: the novelty of the tool for users (lack of familiarity); our aim to include participants from different countries, which created some language barriers (the 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 previous analysis of video captions in WP5, was that video description captions are currently not sufficiently accurate or reliable to provide a realistic starting point for a simulation of real-life editing).

The **outcomes** of the study will increase our understanding of this new human-in-the-loop workflow for content description and inform the further development of the tool. Although this evaluation centered on the appropriateness of the tool for archival/re-use contexts, it has implications for audio description and, more broadly, media accessibility for audiences with visual and cognitive impairments.

9.1 Methods of evaluation and user test setup

The workflow examined in this evaluation is one in which the human remains a crucial element in the loop while automation aims to solve an important precursor role, with the curation of functional video descriptions generated first by AME services and then corrected manually. Given the novelty of this workflow, it was considered appropriate to adopt a mixed-methods approach, involving, 1) observation of participants undertaking hands-on work; 2) a validated instrument to gauge user experience, the UEQ, complemented by questions about participants' experience of their work environment and their preferences with regard to specific features and functions of the prototype; 3) focus groups to elicit participants' views and suggestions for further development. The basis for the data collection was a set of workshops with the participants.

9.1.1 Material used

As material for the content description evaluation, a set of five video clips were selected from a variety of programs provided in the YLE content data sets (cf. Section 5.6), a collage of which is shown in Figure 22. Each clip was between two and three minutes long and they covered the following content:

- A clip from a lifestyle program (i.e., "Strömsö" in this case);
- Two clips from an older current affairs program discussing a visit of England's Queen Elizabeth II to Finland;
- A clip from a recent current affairs program about politics in the USA;
- A clip from an older broadcast discussing the impact on the life of citizens of the UK after joining the EU in 1973.

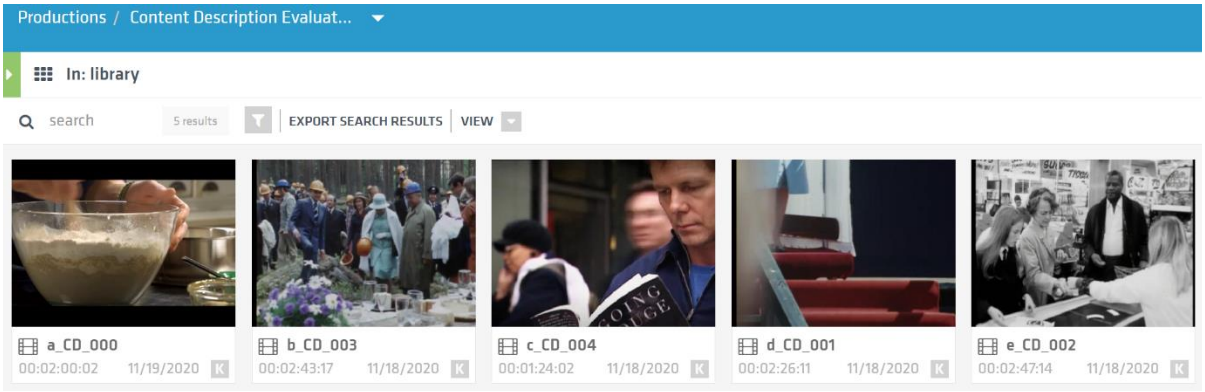


Figure 22: Five video clips selected for the content description prototype evaluation.

9.1.2 Participants

The participants formed a convenience sample (recruitment being conducted through organizations already expressing an interest in the work of the MeMAD consortium, including members of the MeMAD External Collaborators Group) and were drawn from the broadcast and media archive industries including broadcast archives. A total of 23 prospective participants was recruited. Four cancelled their participation before the start of the workshop; one withdrew at the beginning of the hands-on session of the workshop due to technical problems on their side.

Participants completing the workshop (N=18) were employed in roles in television production, the archive department of a broadcast company, and a national film archive. Their job titles included production coordinators, assistant producers, archive journalists and cataloguers.

Participants were recruited in four countries (Finland, Sweden, Switzerland and Germany). Their ages ranged from 30-49 to 60-69, with 67% identified as female and 33% 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 had between 1 and 10+ years' experience in the description of AV content, as part of their work in either archive or broadcast institutions.

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.

9.1.3 Study design, user data collection and tasks

The participants were invited to participate in an **evaluation workshop** in which they were given an induction to the platform and then had the opportunity to pilot the prototype in an individual hands-on session. Embedding the evaluation in a training workshop was thought to be the most effective way for participants to take part in this study, providing them basic familiarity with the tool (but no more than that) while enabling the research team to observe the process and elicit participants' initial views of the platform. Due to the COVID-19 situation, most participants worked from home, using laptops. They were given access to the MeMAD platform through a broadband connection. Zoom video conferencing software was used for the training workshop.

Six workshops were held, each with up to five participants and each following the same pattern. The first part was a 45-minute **induction** to the project and the platform by the MeMAD project team including the software developers, and allowed time for a brief Q&A, due to the small number of participants. This was followed by a 45-minute to one-hour **phase of individual work**, using the Zoom breakout rooms. Participants were given access to five short video clips (one was used as a "warm-up" exercise) and briefed on the required task(s). The speed at which each participant worked varied according to experience and technical competence, with the result that the number of annotation tasks completed ranged from between one and five video clips. Three participants completed four clips, six participants completed three clips, a further six participants completed two clips; three participants worked on just one clip. In addition, participants had access to individual technical support in their Zoom breakout room throughout the hands-on session.

After completion of the hands-on session, a **questionnaire** was administered to elicit the participants' views on and experience of working with the platform. This was followed by a **focus-group** style discussion, which lasted between 33 and 71 minutes across the six workshops.

9.1.3.1 User data collection

To observe the video describers' approach to the description task and their interaction with the prototype, participants were asked to share their screen in Zoom during the hands-on session, and the sessions were video-recorded. The focus groups were also video-recorded using the Zoom platform.

The questionnaire consisted of four parts. In addition to collecting basic demographics, it incorporated the User Experience Questionnaire (UEQ) and two additional sections. The UEQ has been widely used to elicit users' impressions, feelings and attitudes towards a range of interactive products similar to the prototype tool evaluated in this case. As such, the UEQ consists of twenty-six 7-point Likert-type questions, intended to measure usability and user experience across six 'dimensions'. (cf. Table 16).

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?

Table 16: Dimensions of the UEQ.

Although the content description tool is at the prototype stage, we decided to use all components of the UEQ in this evaluation. Whilst we were mainly interested in the usability aspect, it was thought to be useful to elicit participants' views on the attractiveness and user experience to inform future development. However, the prototype nature of the tool has been taken into account when interpreting the outcomes for these components of the UEQ. The full UEQ used was included as Appendix D.

The second part of the questionnaire contained two further sets of 7-point Likert-type questions, one relating to the participants' experience of the **work environment** and another eliciting their preferences regarding specific **features and functions** of the prototype. The work environment section sought to elicit participants' overall impressions of process and workflows (e.g. *"I felt comfortable working in this* *environment”*) and their perception of enhanced 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”*).

9.1.3.2 Data analysis

As the user-centered evaluation of the usability of the prototype application’s design and GUI formed the focus of this study, the UEQ’s features in this area could hence be utilized to their fullest. The UEQ part of the **questionnaire** was analyzed using a tool integral to the UEQ evaluation package. This provides basic quantitative data on user experience by comparing participants’ recorded answers with a benchmark dataset, containing metrics from over 14000 participants evaluating more than 280 products (e.g. business software, web pages, online shops, social networks). The findings section provides benchmarked mean scores for each of the six dimensions of the UEQ. 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 specific to this particular study. Answers to open-ended questions were analyzed qualitatively with a focus on user preferences.

Focus group discussions were analyzed thematically to determine, compare and contrast participants’ perceptions of the prototype. Where 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 aim was to use the focus groups to make a deeper dive into the observed actions and questionnaire responses obtained from the earlier phases of the study.

9.1.3.3 Ethical considerations

The study was approved by the University of Surrey Ethics Committee (Reference number: FASS 20-21 014 EGA). As highlighted above, the study participants were from four different countries. Most participants indicated that they would be comfortable participating in English. However, the questionnaire was made available in English as well as Finnish to accommodate different language backgrounds. For the same reason, some of the focus groups were held in Finnish, others in English.

9.2 Analysis of user data

9.2.1 Findings from the survey – UEQ

As discussed in subsection 9.1.3.1, the twenty-six 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 17; the scores for each of the six dimensions are shown in Table 18 and Figure 23. UEQ scores range between -3 (extremely bad) and +3 (extremely good). 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 generalizability, 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.

It.	Mean	Var.	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 17: UEQ mean scores for individual questions.

Dimension	Mean	Var.	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

Table 18: UEQ scores per dimension.

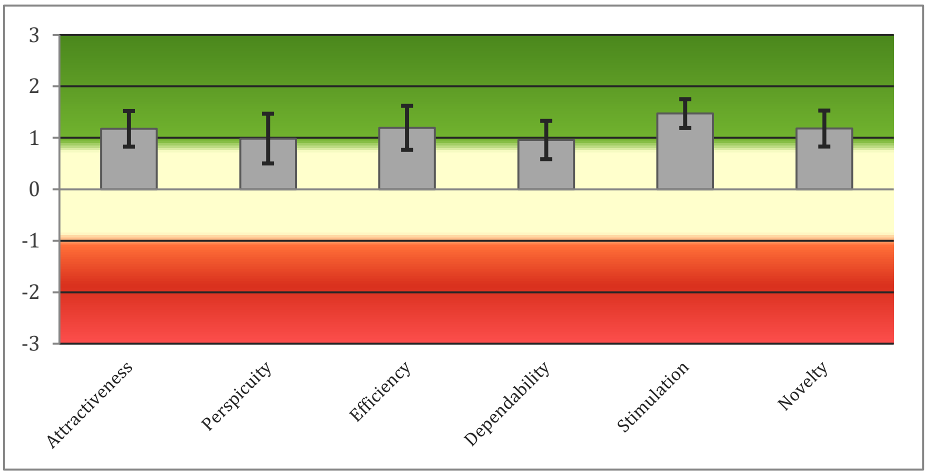


Figure 23: UEQ scores per dimension.

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 the content description application 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 descriptions.

In order to place these scores in the context of industry standards, the study's UEQ results were benchmarked against a 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 the application are shown in Figure 24. 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).

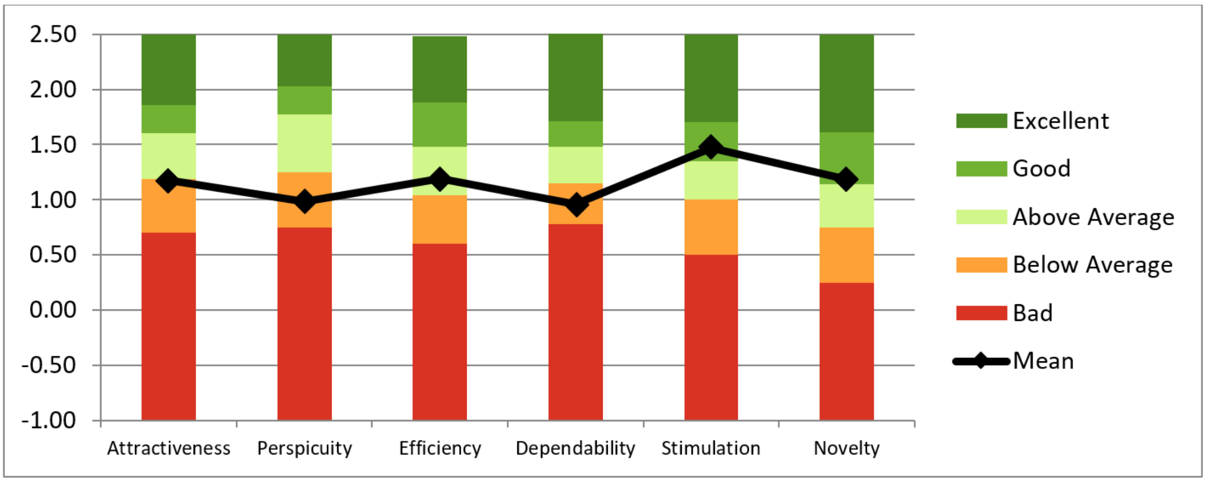


Figure 24: Benchmarked UEQ results.

Figure 24 shows mean scores (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 mean score of *attractiveness* (M=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 the prototype 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.

9.2.2 Findings from the survey – 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 (Figure 25), based on the participants’ perceptions of:

- 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;
- 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).

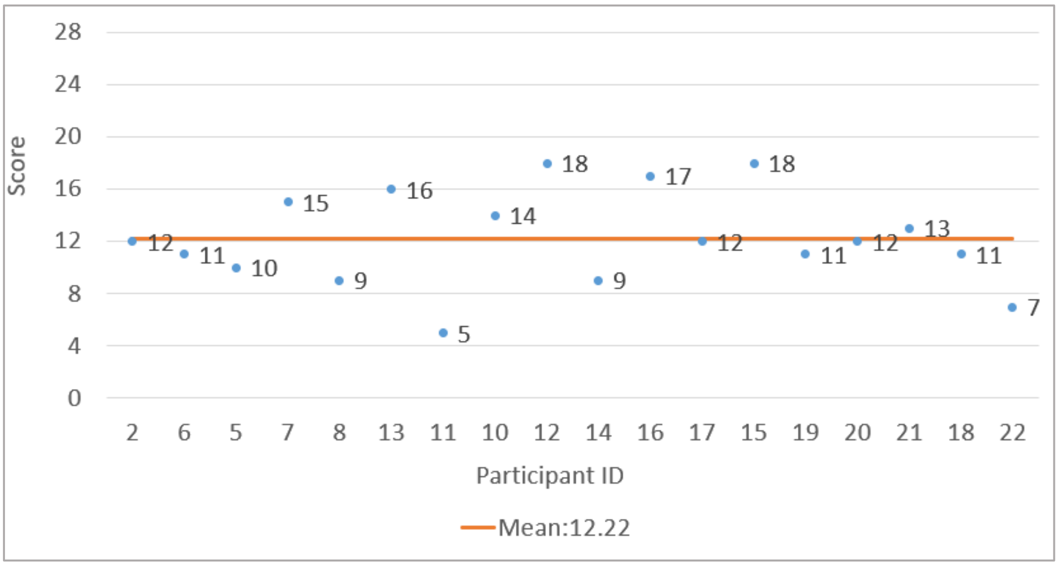


Figure 25: Experience of working environment among the participants.

The most positive perceptions were observed in the group indicating 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 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 company's current workflows but it may in part be explainable by the fact that the organizations 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 pre-segmented 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 issues with 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 (expert) and processing 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. 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.

9.2.3 Findings from the survey – 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 19) were rated average or above.

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) 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

Table 19: Views about specific features of the content description editing prototype.

Whilst 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 in 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 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 on the timeline lanes. The feedback garnered in the focus group discussions (see subsection 9.2.4) 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 (Figure 26).

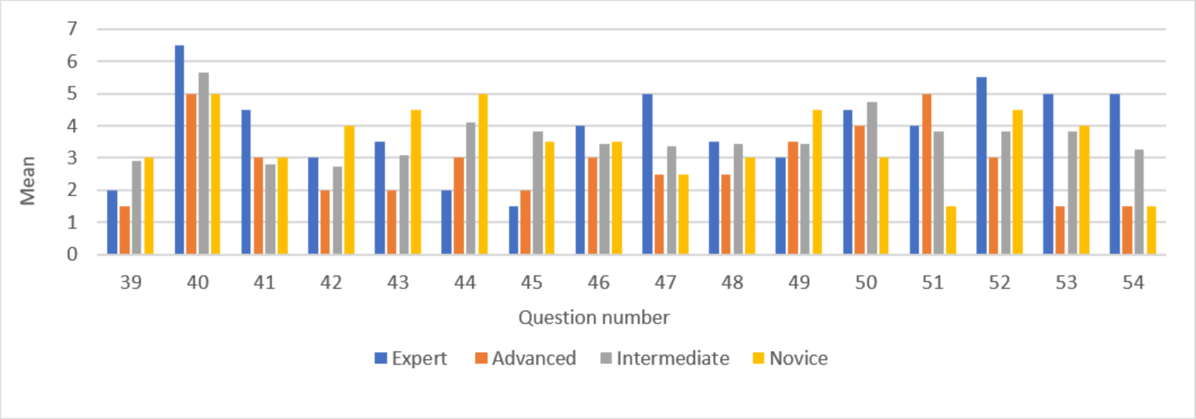


Figure 26: Features and functions according to level of expertise (1/2).

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

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:

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 (Figure 27).

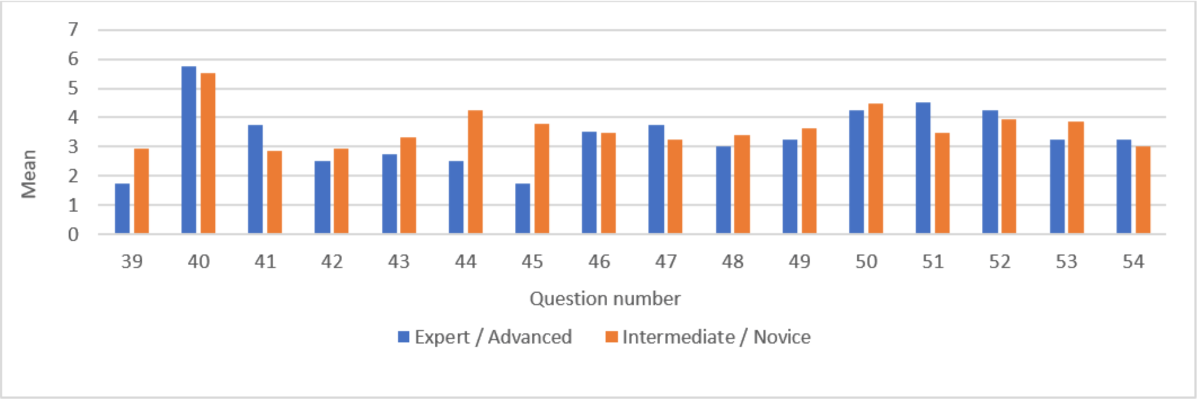


Figure 27: Features and functions according to levels of expertise (2/2).

9.2.4 Findings from the focus group interviews

9.2.4.1 Usability of the prototype - overall perception

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 the prototype was described 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 they 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 prototype application at a glance.

A recurrent theme across the focus groups was **familiarization**. Several participants expressed regret at not having been given more time in the workshop to familiarize 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 responds 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 ones.

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 (a finding which resonates with WP5's recommendation regarding improvements to video description saliency noted in cf. D5.3, section 5.3.5.). 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 in place 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, as well as aspects that could be further improved. These are summarised below.

- The tool was commended for its **overall user-friendliness**, including its clear layout and ease of use (e.g. navigating through the video clips).
- The **presentation of the 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, in the same place, obviating the need to search for metadata and enabling the user to choose and decide whether to use a given (machine-generated) description. One user commented that this increased their motivation to use the automatic metadata.
- 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. Moving through the timeline was different from most editing platforms, although it was easy to learn.
- The use of different **lanes (tiers)** to display the various types of metadata was thought to be helpful. This is corroborated by the ratings given in the features and functions section of the questionnaire. However, some participants reported that they did not make much use of the lanes due to their location at the bottom of the screen, which was different from the tool they generally worked with and difficult to see on a laptop screen. The ability to **type a description while the video clip was running** appealed to participants.
- **Keyboard shortcuts** were highlighted as being helpful.
- The availability of **separate fields** for automatically generated data and the human-made description was deemed to be useful for accuracy and reliability.

The participants also pointed out that **traceability** (i.e. the possibility to see whether the material originated from another clip) is important, in the context of re-use rights. They pointed out that it can be problematic if re-use rights are not clear for a clip that is re-used.

Aspects requiring improvement

- Participants in one of the focus groups felt that the platform should offer more **room to tell a story**, e.g. by incorporating a storyboard function. This is particularly interesting to note, as it aligns with ongoing work in WP5, which suggests that basing this on a story grammar segmentation format could allow users to 'sketch out' the full story arc in advance.
- Some participants queried the helpfulness of entering **free keywords**, suggesting that lists of controlled vocabulary would be more effective.
- Another concern was that **repeated descriptions** were required across contiguous segments for elements such as persons and places, i.e. that the labels had to be typed repeatedly. Participants from one of the companies in the sample explained that their normal practice was to describe places and characters occurring across several segments only once and/or that they can copy and paste the information, if they need to repeat it.
- There was some uncertainty as to whether a description, tag or session had been **saved**, which led to re-entering descriptions repeatedly.
- As was apparent from the specific features and functions section of the questionnaire, the **'set in set out' feature** was difficult to use.

9.2.4.2 Individual 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 summarizes 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 deleted if necessary.
- It would be important to be able to quickly delete inaccurate information or to mark it as inaccurate (which could then be used as training data). This was deemed easy to achieve in the case of individual, inaccurate words.
- 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:
    - Full sentences seem to be difficult to produce automatically. It would therefore be more helpful if the machine gave keywords which I can delete or confirm, and then I write the sentences.

- If you have a guideline according to which you have to write descriptions in full sentences, then it can take time to correct the automated data; it can be easier to write them yourself.
- Automatic suggestions are not useful for content description, but tags may be useful to the end users.
- I would like to have only one method, and this was mixed: persons, suggestions and text. Sometimes you had to delete and sometimes to activate (confirm), which was complicated. (Note: other participants, by contrast, regarded the way in which suggestions are now available as interesting and easy in use).

Segmentation/shots

- Participants who describe shots or similarly short segments in their normal work practice found the segmentation helpful and used it in the evaluation.
- Participants who do not describe content at shot level found the segmentation too detailed.
- None of the companies involved in the evaluation generally describe every shot; the minimum length of a segment for description is normally five seconds.
- The segmentation feature was seen as useful for programs with several sub-topics.
- Participants also emphasised that segmentation helped them move through the clip and to check where they are.
- A functional problem was noted in connection with the segmentation, namely that it took too long to edit the timeline and that this had an impact on segmentation.

Speech transcripts

- 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.
- The transcript was deemed especially useful for the description of talk shows and factual programs and where a celebrity says something noteworthy.
- It enables the describer to check names of persons and places as well as quotes.
- In addition to supporting the creation of content descriptions, the transcript can also support journalists in writing the program script.
- It is furthermore useful in subtitling for the deaf and hard of hearing.
- Regarding the quality of the transcript, the Finnish versions were found to be reasonably good but dependent on the sound quality; the English transcripts were of good quality.

Machine Translation

- Translations were found helpful for the description of foreign-language programs, as a way of gisting.
- Similarly, the translation was found useful for gisting purposes in journalists' work.

- 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 programs and archives.

Face Recognition

- Participants found the idea of automatic face recognition useful but questioned its reliability.
- The feature was deemed to be particularly useful for describing old material which does not have metadata, but it was acknowledged that this material would be challenging for automatic face recognition.
- Another use that was highlighted was the identification of foreign names through face recognition, as this would help with spelling.
- Participants also pointed out that it needs to be possible to correct the spelling of names and that a user's correction should be applied to all occurrences of the same name across the clip.

Key elements, tags and named entities

- 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).

9.2.4.3 Suggestions for improvements

During the discussion of the different functions of the prototype, FG participants also made suggestions for improvements and additional features.

Suggestions for improving existing functionality

- When a user moves back and forth on the timeline, there needs to be a way of indicating whether changes have been saved, or the user needs to be prompted to save changes to prevent data loss. One participant, for example, moved the timeline while still in edit mode for the content description segment they had been working on, without saving the description. This meant that the description was lost. The difference between the 'editing' mode and the 'selecting time range for new content description' mode can be made clearer in the application. Explanations for the fields would be useful, e.g. in the form of 'tool tips' when the mouse/cursor is hovered over the field in question.

- An option to add document-level metadata applying to the whole clip would be useful.
- An option to close some of the fields when searching would be useful.
- Audio capabilities, e.g. a way to toggle between different audio channels could be developed.
- Thumb nail images could help navigate through the video.
- The space on the screen could be used better: in the left corner underneath the video much space was not used, whilst in other areas (e.g. the lanes) the volume of information felt burdensome.

Suggestions for additional functionalities and metadata

- The topic of the segment could be captured; speech does not necessarily contain the topic, and a topic is more abstract than an image. The topic field could also contain a summary of a discussion, e.g. in a talk show.
- A vocabulary or ontology, similar to Wikidata, could be added.
- A facility to note additional features for persons/speakers would be helpful; for example, this might indicate whether a person is speaking, or whether the speaker is visible, or indicate the language in which someone is speaking.
- Similarly, a facility might be added to capture the additional features of an image/shot, e.g. to indicate whether it shows the outside or inside of a building, and to indicate the type of shot (close-up, aerial view, black and white) and format ratio.
- Automatic music recognition would be very useful, as most TV shows have music in the background.
- Automatic recognition of buildings would also be helpful for the descriptions.
- A storyboard would be highly desirable, as it would enable describers to see the entire file at a glance, i.e. an overview of all segments (this links to the work on Moment Detection in Deliverable 3.3).
- A time lane for subtitles should be incorporated.
- A field for authors should be included.
- A confidence estimation for the automatic captions should be included; alternatively, only captions that pass a confidence threshold, e.g. 95% should be included.
- It would be desirable to gather as much information as possible from the creative process, and describers should be able to trace the origin of the materials.

9.3 Discussion of the user feedback and data gathered

The wealth of user feedback gathered throughout the evaluation, by means of the UEQ, the focus group interviews and the screen recordings obtained from the various evaluation workshops permit us to make a set of observations concerning the future development of the prototype and the viability of the proposed human-in-the-loop workflow. We discuss these observations in detail in this section and summarize the main conclusions and next steps in Section 10, along with findings from evaluations on the other Epics implemented by the MeMAD platform.

9.3.1 Considerations for future user-interface and user experience changes

This section extracts concrete improvements to be made to the application from the feedback from the previous sections.

Vertical spacing between various elements

No explicit user feedback was given regarding the spacing of UI elements on the screen. However, from analyzing the screen recordings when looking for the cause of specific user confusion and feedback, we identified two problematic areas in the user interface.

The first problematic area is between the timeline and the top row of the player controls (including the ‘SET IN’ and ‘SET OUT’ buttons), highlighted in Figure 28. We noticed users wanting to click the ‘SET IN’ button who accidentally clicked the green player scrubber bar which is just above it. This indicates the green scrubber bar is too close to these buttons. The same problem also exists at the top of the green scrubber bar: users wanting to click the bottom lane of the timeline sometimes accidentally clicked on the green scrubber bar.

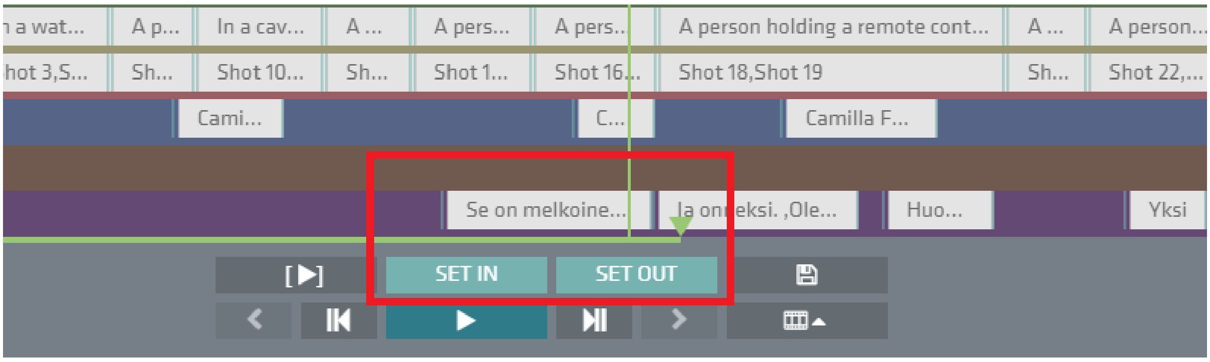
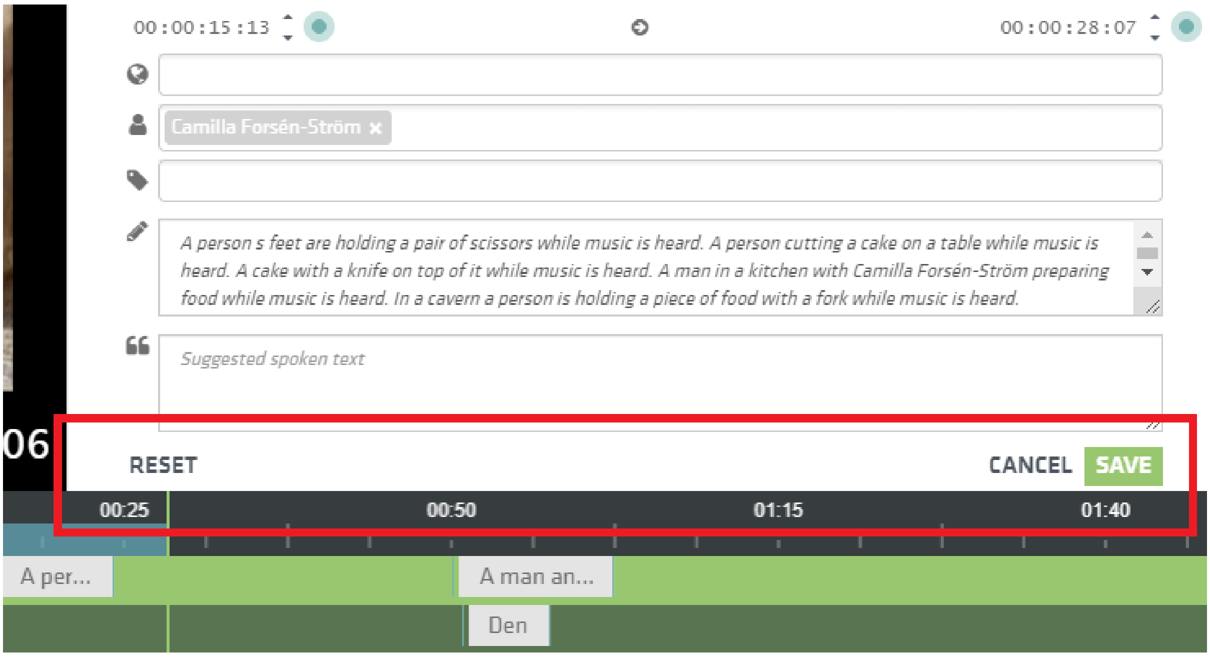


Figure 28: Vertical spacing between timeline and player controls.

The second problem with vertical spacing which was evident from analyzing the screen recordings is the lack of vertical spacing between the ‘Cancel’ / ‘Save’ buttons and the timeline below them (Figure 29). This caused some users to accidentally click the timeline, which changed the time range they had selected, just before they tried to save the content description.



00:00:15:13

00:00:28:07

Camilla Forsén-Ström ×

A person s feet are holding a pair of scissors while music is heard. A person cutting a cake on a table while music is heard. A cake with a knife on top of it while music is heard. A man in a kitchen with Camilla Forsén-Ström preparing food while music is heard. In a cavern a person is holding a piece of food with a fork while music is heard.

Suggested spoken text

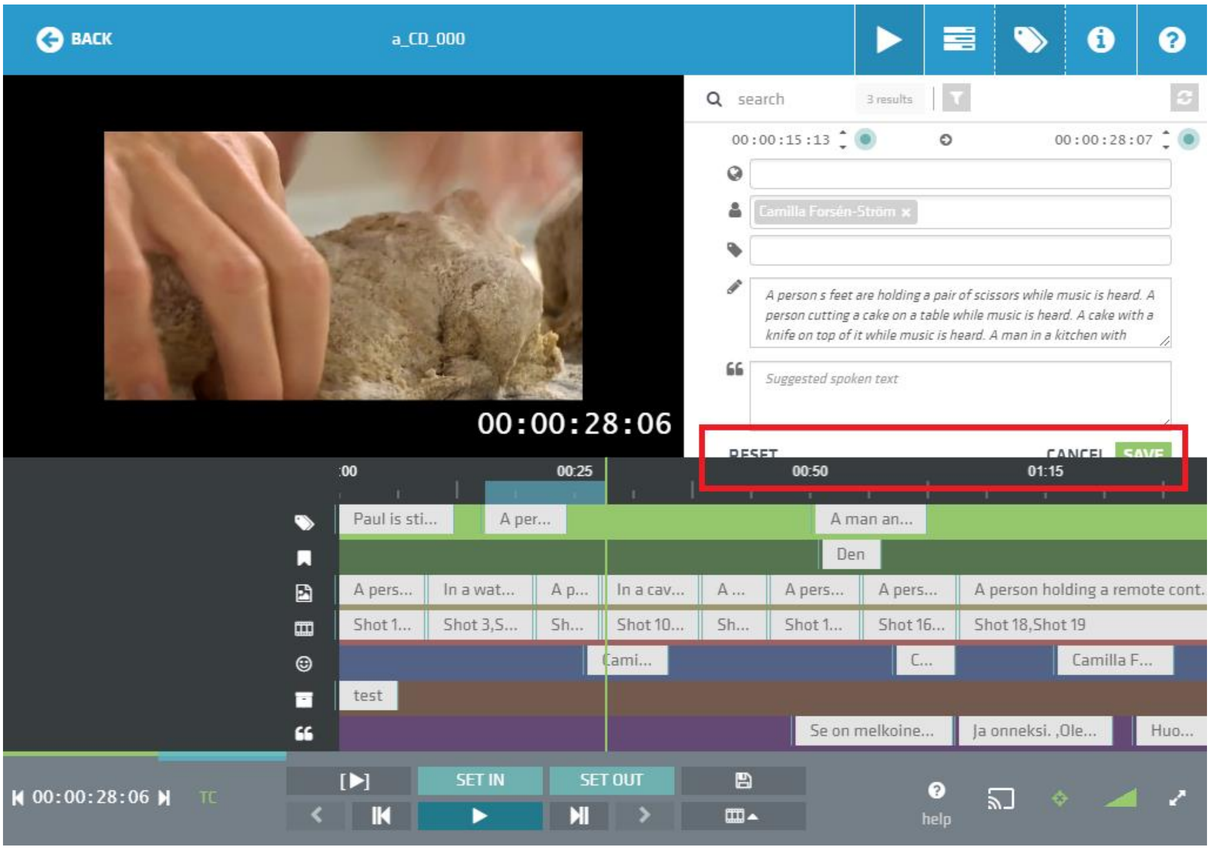
Figure 29: Vertical spacing between the timeline and Cancel/Save buttons could become problematic due to the proximity of the action buttons and the interactive timeline..

Add support for smaller screens

We overestimated the resolution of the participant's screens. The designs were made to work with a resolution of 1280x800 or above, but many participants worked on a machine with a lower resolution. Especially a limited vertical size (smaller than 800px) caused problems, as the header, player controls and timeline all take up some size which makes the remaining size for user input elements too small.

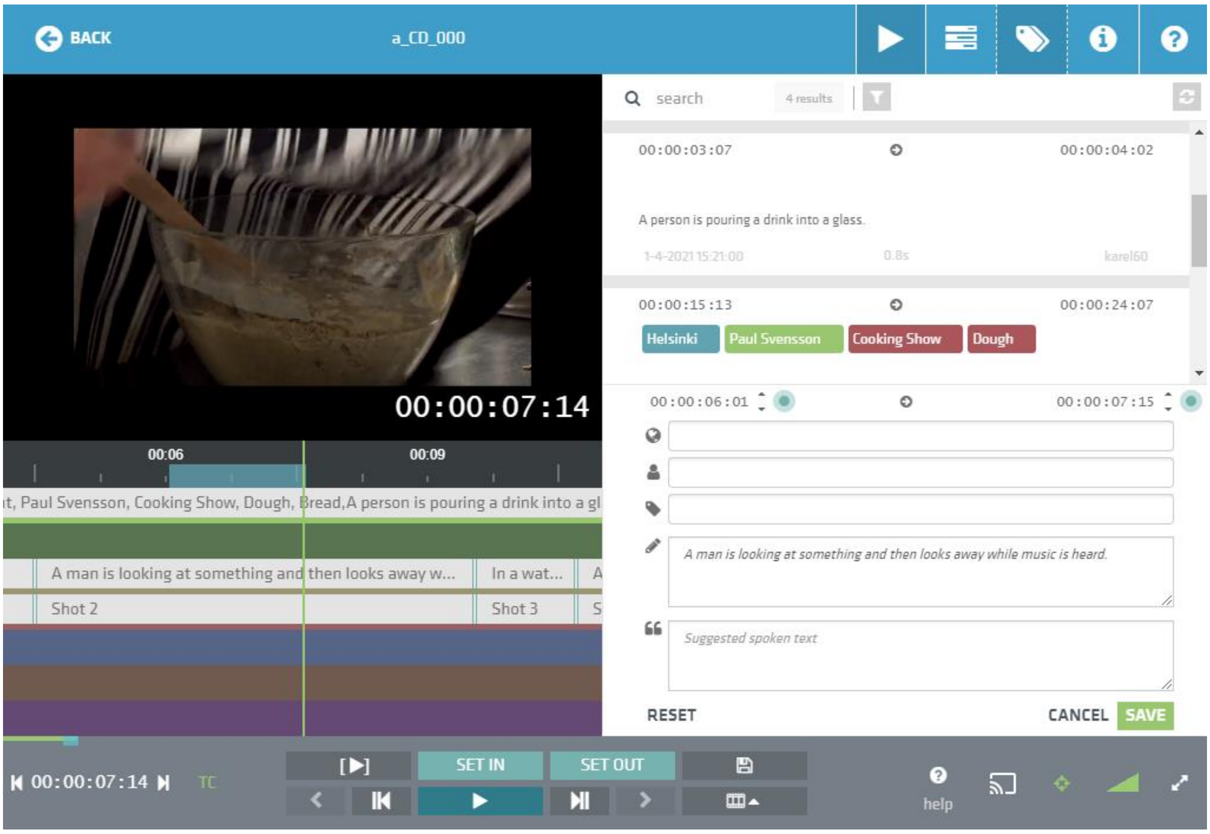
Figure 31 illustrates the problem when the vertical size of the application window gets too small (the screenshot is at a resolution of 1024x715). The 'Reset', 'Cancel' and 'Save' buttons start to slide below the timeline, becoming completely invisible when less than 700px of vertical height is available.

Following a suggestion by the study participants to better use the space on the screen (see paragraph 9.2.4.3), a solution to this problem could be to not extend the timeline over the full width of the screen, as shown in Figure 30.



A person s feet are holding a pair of scissors while music is heard. A person cutting a cake on a table while music is heard. A cake with a knife on top of it while music is heard. A man in a kitchen with

Figure 31: Problems with function buttons becoming invisible due to small screen size.



A person is pouring a drink into a glass.

1-4-2021 15:21:00 0.8s karel60

A man is looking at something and then looks away while music is heard.

Figure 30: Solution for problem with small screen size.

Metadata timeline optimizations

During the design of the application (cf. D5.4), the decision was made to always fix the player position to the center of the timeline (indicated by the vertical green line, shown in Figure 32), such that the timeline effectively scrolls underneath this line as the content plays. The reasoning was that this produced a more consistent view because the player position can always be found at the same screen location.

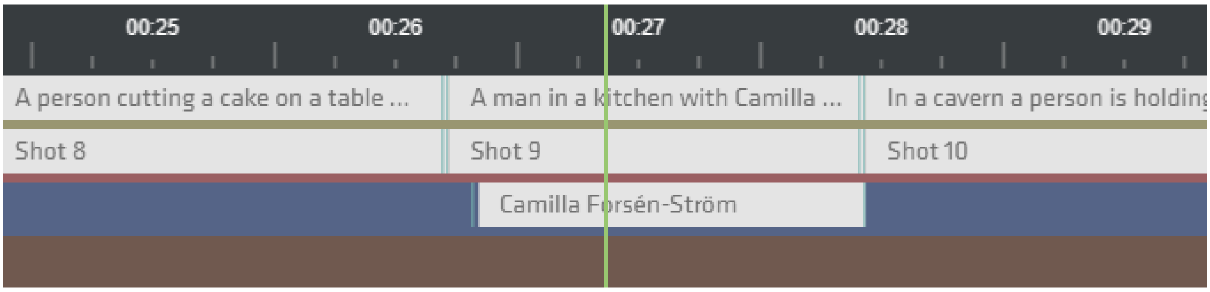

00:25
00:26
00:27
00:28
00:29
A person cutting a cake on a table ...
A man in a kitchen with Camilla ...
In a cavern a person is holding
Shot 8
Shot 9
Shot 10
Camilla Forsén-Ström

Figure 32: Timeline with player position fixed to center.

However, we failed to recognize that this is different from existing non-linear editing software the users might be familiar with. As a result, some users were confused at first and tried to change the player location by dragging the vertical green line instead of dragging the timeline. Most participants found it easy to learn though. In hindsight, we should have stressed this change more during the induction phase or made this more clear in the user interface.

Time range selection

Creating a new content description starts with selecting the appropriate time range. The application provided two methods: either by clicking the ‘Set In’ / ‘Set Out’ buttons or by clicking a segment on the timeline.

During the design phase of the application (again, as discussed in D5.4) the YLE archivists indicated that the method using ‘Set In’ / ‘Set Out’ was familiar from current tools. Interestingly, four participants in the evaluation study indicated that they found the ‘Set In’ / ‘Set Out’ buttons difficult to use. Further analysis of the screen recordings revealed that for two of them, the problem was actually caused by accidentally clicking the player progress bar while intending to click one of the buttons, so they are, in fact, instances of the problem were already discussed above.

For the remaining participants, the problem was caused by the use of the ‘Set In’ button while still editing the previous content description. This did not create a new content description but rather changed the timing of the content description that was being edited. A better visual indication that the application is in the ‘editing’ state together with contextual tooltips on the ‘Set In’ / ‘Set Out’ buttons would improve this.

The other method to select a time range, clicking segments on the timeline, was not used as much as anticipated. The participants indicated that the shot level is too detailed (cf.

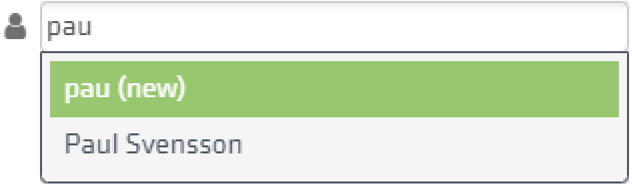
subsection 9.2.4.2 - ‘Segmentation/shots’). It seems it was not clear to them that multiple shots could be selected at once by shift+clicking them. An improvement would be to make this more clear, for instance by showing a tooltip while hovering the mouse over the segments.

Manipulation of the input fields

The ‘opt-in’ behavior of the key elements (people, locations, other tags) was preferred over an ‘opt-out’ behavior by the YLE archivists during the design phase of the application. This did indeed prove to be appreciated by the participants in the evaluation.

However, the ‘description’ and ‘spoken text’ fields also display suggestions based on the DeepCaption and ASR services. These fields are currently ‘opt-out’, in that the user has to explicitly clear or overwrite these fields if the suggestion is incorrect. The combination of ‘opt-out’ and ‘opt-in’ fields was found confusing (cf. subsection 9.2.4.2). An improvement would be to make these two fields also ‘opt-in’, requiring the user to accept them before applying them on the content description.

As indicated in section 9.2.4.1, some participants expected the input fields for the key elements to be linked to a controlled vocabulary, which ensures consistent metadata and avoids spelling mistakes. For the proof of concept, the input fields accepted any text (as illustrated in Figure 33). We did however suggest terms that were added previously on the active clip. To limit the development effort for the proof of concept, we did not integrate a thesaurus, but the application design could support this in the same way.



pau

Figure 33: Working with suggested terms, in free-text fashion.

9.3.2 Semi-automating AV content description as a novel workflow

On the face of it, this evaluation was aimed at the prototype content description editor, 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 the 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 headline 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 usable 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 work on explicating the human process of discourse comprehension and production (WP5), which has demonstrated that this process is largely holistic, involving continuous attempts at creating a mental representation of the story emerging from any given (verbal or multimodal) text, using cues from this text to activate common knowledge which, in turn, helps us to integrate the elements in a discourse to create 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 (cf. Deliverable D5.3). 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 contextualizing 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 contextualization 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 **customization** to cater for different purposes and settings. Customizable options would also cater for **personalization** to accommodate different working practices and styles, as well as the user's own preferences.

Ultimately, customization 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 additional technical requirements (e.g. fitting the description in silent moments in the audio track).

9.3.3 Limitations of the study

9.3.3.1 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 (see, for example, the issue with saving content descriptions, outlined in section 4.3). Such problems were to be expected, given that the participants only learned how to use the prototype during the workshop.

9.3.3.2 Confounding factors

Whilst the evaluation was focused on the usability and functionality of the prototype, many participants raised related topics that were considered in other MeMAD workstreams, especially the issue of accuracy and reliability of the automatically generated metadata and content descriptions (as discussed in WP5). It is possible that the quality of the sample video captions and person/location 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.

Furthermore, as all the sessions 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. Whilst these are unlikely to occur in a professional work environment, they are worth noting for workplace arrangements to be made in the post- COVID -19 recovery phase, and that they may have affected the participants' perception of the tool at the time of the evaluation.

9.4 Conclusions

The main aim of this evaluation was to find out to what extent, in the perception of professional content describers, the application prototype supports the creation of functional video descriptions and the way individual features and functions of the platform are perceived by professional describers. Through a mixed method approach we have been able to measure the perceptions of the study participants in relation to key dimensions of user experience and usability, and key features of the prototype, as well as capturing qualitative feedback on the human-machine workflow, individual functions of the prototype and suggestions for further development.

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, i.e. a novel technology-enhanced human workflow, whereby human content describers draw on and post-edit machine-generated metadata and video captions to create AV content descriptions, this was acknowledged by the participants as a useful way forward. Whilst the main caveat was, as expected, the quality that is currently achievable for machine-generated material, the participants highlighted a number of benefits of the workflow under evaluation. For example, the automated captions were thought to be useful in identifying aspects of the content that a human operator might overlook, in achieving greater consistency of description and, more broadly, in bringing legacy material without pre-existing metadata into the fold of digitally searchable broadcast/media archives.

Whether this workflow will save time and/or improve the quality of the descriptions was discussed but not measured in this evaluation, as it was expected that the low quality of the automated data would currently skew the results of such an evaluation. A combination of automated workflows and human post-editing has become common in monomodal translation (i.e. translation of verbal text) and *multi*modal translation (i.e. subtitles and other translations of verbal text including images), and has been shown to produce improved results especially with the arrival of machine translation, although differences between language pairs remain strong. However, to the best of our knowledge, the present evaluation is the first of its kind – i.e. the evaluation of a prototype tool that supports post-editing of different strands of automatically generated textual content (incl. metadata, narrative), produced in an automated process of *intermodal* translation (images to text).

10 Discussion and impact of the final prototype evaluations

In this section, we summarize the conclusions from the entirety of end-user evaluations of our prototype platform undertaken in the project. While the previous sections already provided some discussions and insights based on the reported user data and feedback, we dedicate this section to a broader look at the conclusions from each of the evaluation studies. First, we provide a general conclusion for each automated metadata and feature extraction technology evaluated in the project prototype platform, which we follow up with conclusions for each evaluated set of functional user stories that use these technologies. As we can no longer follow up on the evaluations within the project after delivering this report, we do suggest potential improvements and future work.

10.1 Observations from the MeMAD evaluations regarding multi-modal feature extraction AME technologies

The various multi-modal feature extraction AME services built in WP2, WP3 and WP4 of the project were vital for the successful implementation of the functional epics of features devised in the specification of our MeMAD prototype. Even though the prototype we built was evaluated in terms of higher-level workflows in this work package, blurring to some extent the providence of certain elements of enrichment metadata, we did want to dedicate a section to the appraisal of each integrated technology. This assessment reflects the level of maturity we judge each technology has reached given the advancement of the state of the art thanks to this project and the usefulness of the technology in the media production workflow. At the same time, we also wanted to highlight remaining challenges that still hinder a wider adoption in some cases.
Note that more services were developed in MeMAD than we addressed here. We included those for which a clear perspective was attained from the point of the view of the prototype evaluations and proof-of-concept projects (cf. D7.4).

10.1.1 Automated Speech Recognition (ASR)

Relevant evaluations: *all.*

We found ASR to have become a mature technology for many cases of media production. It can be employed in a variety of use cases, including for content retrieval, auto-subtitling, and as an aid to content description, and often serves as a precursor for other processes such as NER or machine translation. ASR works best of all for direct use of the produced text transcripts. Second order errors can have a big impact on overall usability, cf. also our assessment of machine translation, which is a point of attention considering that ASR is the prime way of providing insight into audio signals and many other metadata is derived from the ASR output.

Introducing a limited post-editing correction step can help mitigate this problem when the most glaring ASR errors are removed before further processing occurs. This was demonstrated in our PoC with France Télévisions on the local elections of 2020 (cf. D7.4); providing a correction interface helps improve the accuracy of later data mining on the speech transcripts.

Accuracy of the ASR output is dependent on clarity of the speech signal and its correspondence to the trained speech models. E.g., dialects or bad pronunciation naturally have an impact on the accuracy of the transcription result. Note that in many cases custom dictionaries can be doped into the ASR process to reduce error rates for domain-specific kinds of speech (e.g., names in sports or domain-specific terminology for medical purposes). MeMAD has demonstrated that there is still room for improvement, as demonstrated by Lingsoft's and AALTO's research on this topic. The project has delivered better accuracy on Finnish and Swedish ASR on all domains making it now feasible for production use.

The maturity of ASR is demonstrated by the large market of ASR solutions available to date, and the fact that, thanks to the evaluations and PoC, and developments of the MeMAD project, YLE has decided to adopt ASR into their intralingual subtitling production on a regular basis. Additionally, ASR already forms a cornerstone of the Limecraft Flow platform today, as we integrate with various ASR service providers and speech transcripts are used either directly for content retrieval and enrichment, and to bootstrap clients' subtitling efforts.

10.1.2 Named Entity Recognition (NER)

Relevant evaluations: Content retrieval (Epic 6.3) and video editing (Epic 6.5).

As with ASR, we found NER to have become a mature technology. Depending on the accuracy of the source text given, NER processes were used in many cases using speech transcript inputs. NER outputs and disambiguation with e.g., multi-lingual labels for named entities is a useful source for enabling cross-lingual retrieval. Naturally, the better the NER tool understands the context of the text provided, the better the disambiguation of terms in the result will be.

We observed from the Epic 6.3 evaluations that NER results were not perceived by users to be directly usable; they saw little direct benefit for it because search results were already returned based on textual queries that matched in the ASR output transcript. However, they did also ask for consolidation of concepts and relevant words from transcripts into broader categories or according to a thesaurus or well-defined ontologies. For this, NER will in fact be a crucial step to be introduced between the ASR output and the mapping of terms to their broader category. At the same time, NER results were used indirectly for content retrieval as multi-lingual labels for detected named entities allowed users to find content across more languages than were explicitly implemented in the ASR+MT transcript processing pipeline (cf. D6.5).

The basis laid in MeMAD for NER support in the prototype platform can easily be expanded to serve a more permanent purpose in the Limecraft Flow offering. We can continue tweaking our interfaces based on the GUIs developed as part of the 2nd platform iteration, or even better, using the optimized interface of the content description viewing and authoring application (cf. Section 9 and 10.2.4). The challenge will be to determine how the potential breadth of NER results will be presented to users. We can depict each individual element (as implemented in the prototype), or rather present a summary of most commonly detected entities, or something dynamic in-between which scales with the granularity being viewed at any given time.

10.1.3 Face Recognition

Relevant evaluations: *all.*

Face recognition is fast maturing technology, naturally with a great potential in the media production chain, for content retrieval and description purposes alike directly, but also in the background, for improving outputs obtained from various AME processes. Assigning names to visual content is vital in providing insights into the video content, and even so more than performing the same kind of operation on aural data (i.e., speaker recognition, which we did not address in MeMAD).

Privacy concerns are not to be ignored with this kind of technology, making it crucial that trained models for face recognition can be compartmentalized to ensure large-scale recognition of persons is done with the proper consent of the person in question if relevant. As an example, people participating in a talent show should only have their faces recognized in content for that particular program with which they have a contract, and not beyond. Within this context, though, face recognition of a set of relevant people can be very beneficial for the production crew, providing auto-triage of imagery specific to that production project.

As such, the integration of the EURECOM FaceRec service has proven insightful as it allowed us to self-train the underlying classification model and observe the effort involved in this process. Care should be taken though to ensure the face recognition model works accurately enough when a large set of faces is added to the model. False positives between similar faces could have a severely negative impact if conclusions are ever made from people wrongly associated with video content in which they never appeared. On the other hand, face recognition is a crucial process in uplifting AME metadata to useful levels in the visual domain. As such, it was one of the main goals of the visual caption efforts done in WP2 (cf. D2.3). This was also demonstrated in D7.4, in the proof-of-concept (and ongoing follow-up project) with the Associated Press. Video captions only become truly useful if person names can be attached to actions described by automated systems in a breaking news context.

Challenges remain in improving the accuracy of face detections, and in making face recognition more robust in the light of video content as opposed to still images. People moving in the image, with varying angles of exposure tend to still throw off the face recognition too much and prohibit a clean and continuous person detection throughout the relevant sections of video. Face recognition (perhaps combined with tools for action and pose detection) should ideally be able to provide a single and clean stream of continuous occurrences with spatial coordinates that vary over time of a person's presence in a video content item. Face recognition services should also adopt a feedback mechanism such that misdetections of people can be reported and be avoided at later stages.

The incorporation of some form of face recognition service in the commercial outcome of MeMAD is likely in the short term, either using a derivation of the EURECOM software, a readily available software-as-a-service, or a combination of both. We can envision setting up pilot productions in which a controlled set of content is catalogued automatically per detected person, using a custom-trained model or by employing labeling of unnamed detected faces. If successful, a subsequent step would be indexation on a larger scale, likely using larger databases of public figures for face recognition to serve bigger volumes of audiovisual content enrichment.

10.1.4 Machine Translation (MT)

Relevant evaluations: *all.*

We have found machine translation (MT) to have become quite a mature technology provided that the underlying models were trained on a sufficiently large data set. MT shows viable potential for use cases such as content retrieval in which case we translated original transcripts to a set of query languages, or when adopted for translating existing intralingual subtitles (cf. D4.4). This is a case in which MT already works very well, with sensible and human-curated input, similar to using, e.g., manually corrected transcripts.

The applicability of current MT starts to break down when second-order errors become more prevalent due to chaining of MT at the end of an ASR process. Errors introduced by the ASR process lead to further and larger confusion when transcribed texts exit the MT step. This was clearly demonstrated in the reception of interlingual subtitling automation (cf. Sections 8.2 and 8.3) which suffered from this effect. The understandability of ASR+MT-generated subtitles was greatly impacted by errors that propagated from the ASR process. Hopefully, these errors can be mitigated by building ASR and MT tools that hook onto each better than they do today, e.g., by passing transcript alternatives between the two processes from which the MT could derive better translations. Alternatively, the introduction of a mature end-to-end spoken language translation model can also offer solutions to mitigate this problem.

Thanks to recent advancements in the performance of MT, Limecraft has adopted MT in its standard platform over the course of the MeMAD project. Its first application has been to support interlingual subtitling, but this will likely be extended in the near future with applications for cross-lingual content retrieval (as prototyped and discussed in D4.4).

10.1.5 Auto-subtitling

Evaluations: Intra- and interlingual subtitling (Epic 6.11)

An assessment of the algorithms behind the automated generation of intralingual and interlingual subtitles is discussed in subsection 10.2.3 as part of the conclusions from Epic 6.11 evaluations, as it only but entirely applies in that context.

10.1.6 Visual Captioning

Relevant evaluations: Auto-generation and correction of content descriptions (Epic 6.10).

Visual content captioning, as it was implemented in MeMAD represents technology with great potential, but which is still too immature for broad use. Visual captioning (i.e., based on the DeepCaption toolbox, cf. D6.8 and D2.3) was the most elaborate example of multi-modal content processing implemented in the MeMAD project. DeepCaption combines the baseline captioning technology with face recognition, clues on shot boundaries and speech characteristics to produce relevant captions in which actions and named persons (who perform those actions) are delivered to content retrieval and description processes. As such, its goal is to form an automated basis for content description curation in the prototype application we built in WP5 (cf. D5.4 and Section 9). Visual captioning has made strides to bring better captions, but as noted in D5.2 and D5.3, the way yet to go remains long. The produced visual captions provide factual descriptions of the content analyzed but a more profound sense of context is not present as of the current implementation. There is no awareness of the story grammar involved in the imagery and of higher-level concepts that may play out in the content are not yet being grasped by the DeepCaption software pipeline.

The challenge remains significant; producing such higher-level content descriptions requires a multi-disciplinary approach in which even more different modalities are to be combined than is the case today. Accurate action and pose estimation could help, as would a manual input on the structure of the content being analyzed. While the reach toward completely non-supervised captioning – in the end – is laudable, a lot could be gained by feeding the captioning process with domain-specific input, incl. which persons to expect in a given content item or by providing context concerning the story structure or topics addressed.

The temporal nature of audiovisual content also presents challenges to captioning technology, which should adopt a better ‘memory’ of actions that occur in a program such that repetitions or cause-and-effect of one action into the next could be better understood and lead to more relevant captions.

10.1.7 Optical Character Recognition (OCR)

Relevant evaluations:
Content retrieval (Epic 6.3) and content description authoring (Epic 6.10).

With regards to the use of OCR in the MeMAD prototype, it can be considered a mature and generally usable technology. The application of OCR into the platform was a matter of integrating off-the-shelf services which were largely hands-off in their invocation. Complications such as language detection are already included, and these services return rich structured detection results.
Employing these results as part of the content retrieval and indexing process, however, requires some very specific post-processing to filter unwanted entries and to ensure the temporal character of video content is taken into account (as explained in D6.8). On the

other hand, OCR provides clear insights into content that would be obvious when being watched directly, but completely opaque in an archive system without descriptive metadata. OCR can be used, e.g., to understand burnt-in subtitles, on-screen graphics or other vital textual elements visible in the image. Hence, the use of this technology is especially relevant in the context of archive material.

Considering that the incorporation of OCR services in the MeMAD prototype came about following an explicit request during the preparation of a WP7 proof-of-concept demonstrator (cf. D7.4, Section 4), its adoption will likely be swift. Demonstrations of this integration are requested on a regular basis, and we expect this will lead to the incorporation of OCR in standard product offerings derived from the MeMAD project soon. Depending on the specific use case, we do expect further tweaks to be required to the OCR post-processing currently implemented as a first trial version.

10.1.8 Language Identification (LID)

Relevant evaluations: N/A.

Language identification is a promising technology, shown to work in controlled circumstances, but needs to gain maturity before it can be used in production contexts.

The identification of isolated language segments in audio content with the aim to steer subsequent ASR processes is a valuable one as it significantly increases the usability of unsupervised ASR in multi-lingual contexts. A mixture of spoken languages trips up traditional ASR systems that require a language to be selected before the ASR process commences. Sections in languages other than the one chosen beforehand are run through the same language model, resulting in gibberish transcripts at best or seemingly sensible but completely false transcriptions in worst case scenarios.

The MeMAD project delivered a fully functional pipeline and LID toolkit to implement multi-lingual ASR based on language segmentation (cf. D2.3 and D6.8). We did find that work remains to be done on cleaning up LID detection results such that better clustering can be performed which leads to a better segmentation result. Quickly extending the LID pipeline's classification model with additional language remains a challenge that needs to be tackled before production use becomes feasible. Because the LID performance decreases as the set of classified languages increases in size, it will be crucial to easily select and deploy a limited set of detected languages depending on the context of use.

If these obstacles can be removed, the adoption of LID can make a large difference in reducing supervision of the ASR process in multilingual environments. Current implementations require users to specify a language choice before triggering the ASR processing and often only support transcription into a single language per content item. This would be optimized such that the system would pick the optimum language to use for any given audio segment autonomously.

10.1.9 Content Segmentation

Relevant evaluations: Content segmentation proof-of-concept with YLE (cf. D7.4).

As with visual captioning, the content segmentation pipeline built as part of the MeMAD prototype did not yet deliver a mature implementation ready for broad use. The pipeline itself is fully functional, but the segmentations it produced proved too flaky to be globally usable. It must be said that the content with which the software was tested was not the easiest to segment. Both programs (Strömsö and Urheiluruutu) broadly discuss similar topics (lifestyle and sports) over their entire duration, making it harder for a clear pattern of segments to materialize (cf. D7.4, Section 7). This was made even more difficult in the case of Strömsö which often features a very similar color palette throughout an episode. Later experiments with news content from INA showed better results (cf. D3.3). More work will be required to find the correct parameters and algorithms tweaks to guide both the textual and visual segmentation to more accurate results.

A reliable content segmentation would provide an important first step for many applications. Segmentation can help storytelling functionality and more relevant content retrieval or description (e.g., content is segmented in coherent pieces of the same topic, making them easier to work with). Starting from the segmentation further added value could be created, e.g., by producing automated summaries for each detected segment (in fact, a feature implicitly expected by the participants of the content segmentation proof-of-concept) or by ranking the memorability and reusability of each segment (cf. D3.3).

10.2 Observations from the MeMAD evaluations regarding implemented prototype functionality and workflows

In the previous subsection, we discussed the maturity and use of each individual feature extraction and AME technology. In this subsection, we take a broader look at the media production processes that were evaluated for each functional epic (i.e., in Sections 6 through 9) and present our conclusions on what the future looks like for each case.

10.2.1 Overall conclusions concerning the evaluation of "Editing assistance using multi-modal and multi-lingual metadata (Epic 6.5)"

This second evaluation confirmed many findings from the first round. Participants were enthusiastic about the new technologies being evaluated and found the available metadata useful, but at the same time they found only limited use for it in their daily work. This remained so despite the fact that the evaluation panel and the testing environment were adapted to better suit the envisioned use case. Even in the news production context, practical use of the provided metadata inputs was limited, as extensive use of this metadata would have occurred in the phase before editing, namely the search and content retrieval process, which was the subject of Epic 6.3 and its evaluations. Once this same metadata is loaded into the video editing environment, the limited user interfaces available in contemporary editing software works against effective use of the metadata instead of promoting it. Overloading the material with a heap of AME metadata works counter-productive in many cases and is often beside the point for regular editing duties.
The improvements made to the metadata exchange with NLEs for the final evaluation helped to a limited extent to reduce the metadata overload encountered in the first setup, but the end situation was still far from ideal to spin the assessment around completely.

There is more promise in a very targeted approach to bring select metadata into the video editing environment, as demonstrated by a positive reception of machine-translated transcripts for the purpose of foreign content understanding. Again, though, care should be taken not to overload the editing software with loads of data that it cannot properly display to the end user.
Another possible targeted application hinted at during the last evaluation round is the application of AME technologies on live recordings, for example, with international news feeds that are continuous program streams in which news topics are run sequentially. This concerns never before indexed or described content in which finding a specific scene or object can be very time-consuming but very urgent at times. Providing a trimmed down metadata stream along with the video content in video editing could provide the only and hence crucial information source on content being processed.

Impact on future development of the MeMAD prototype platform

Future developments on this epic's functionality will likely focus on delivering very targeted metadata into NLEs, geared specifically for editing. E.g., shot type analysis, etc. Ideally with only one or a few fields per clip to limit overload.

We will also investigate other editing systems beyond Avid Media Composer and a stock Adobe Premiere installation. An attempt could be made to implement a custom Adobe 'panel' in which a separated user interface can be built that would implement a custom metadata view outside of the limited interface elements available in the standard GUI of these applications. Even then, the feasibility of this development would still depend on the fluidness with which this feature would co-exist in the application. A seamless integration that allows users to work through this data at full speed and with the native tooling (e.g., keyboard shortcuts, tight API-wise integrations) would be essential to make this work in practice.

A follow-up project between Limecraft and the Associated Press (cf. D7.4) is currently ongoing in which extended metadata exchange with Adobe Premier is being investigated to bring speech transcripts and descriptive shot-lists into the editing environment on top of prepared content and live feeds freshly recorded in a news context. This implementation will incorporate lessons learnt from MeMAD's Epic 6.5 evaluations.

10.2.2 Overall conclusions concerning the evaluation of "Searching and browsing for ingested and archived content"

The second evaluation for content retrieval (Epic 6.3) showed an improved assessment of the search system due to more extensive available visually-oriented search metadata. In particular face recognition results and visual scene classification helped users complete search tasks more efficiently. Especially combined with metadata deduced from the aural content (i.e., ASR, NER, MT) which was available earlier a reasonably complete coverage of the content's properties is obtained that allowed the user panel to complete their search tasks successfully. For factual content, this has proven to work quite well, and we found that multi-modal AME can begin acting as a substitute for human curated annotations, provided that the retrieval process can be made to work with very literal descriptions of the content.

This was exemplified by the test panel, as their conclusions were still divided on the efficiency of AME metadata for the content retrieval process. On one hand, users feel that much of this metadata (or combinations thereof) is helpful in finding content given that they somewhat rethink the retrieval process in such a way that they query for exact terms, e.g., from quotes or persons identified in the imagery. In these cases, the system returns very precise search results which was appreciated by the test panels. This situation naturally results from the technologies employed for describing content: literal transcriptions of speech or visual observations are currently on offer, but significantly broader interpretation is performed on this base metadata. This was also the observation in WP5 (cf. D5.2 and D5.3), where the "key elements" and basic forms of content description were found to be well represented in the AME output but semantically richer interpretations in form of narrative grammar understanding, event narration and coherent understanding of actions depicted in the analyzed content are still absent. This is exactly the area where human content annotators have their place, as was also found in the evaluation of Epic 6.3. Users felt the metadata was mostly complementary to manually curated content descriptions added by professional archivists. This was due to the ingrained methodologies followed by the test panel, e.g., by searching for generalized terms instead of direct quotes as an archivist would enter them. The question was asked whether the use of controlled vocabularies, a thesaurus or ontology could be adopted to help this search approach. We feel this should be the next step to improve the search experience. The use of a thesaurus or ontology, with terms derived from the AME metadata would help describe items as people know them and would present a further step to automating content description.

Generalization technologies such as topic detection or text summarization were not explicitly implemented in MeMAD but could help implement this process in practice. These services can facilitate in transliterating literal terms obtained from aural and visual descriptors into more commonly used topics. In fact, independent testing by YLE has shown that auto-tagging ASR outputs using these services has proven to be quite tolerant on individual ASR errors (as opposed to other 2nd tier technologies such as NER that suffer from error propagation, as discussed in subsection 10.1.1). At the same time, they can be helpful too in reducing the amount of metadata automatically generated, to mitigate the "too much metadata" concern voiced in by users in the evaluation panels.

Conclusions on the usefulness of each type of metadata produced as part of the content retrieval enrichment process is given in the previous section 10.1.

Impact on future development of the MeMAD prototype platform

The fact that the test panels were unfamiliar with the Limecraft Flow system that was used as the basis for the MeMAD prototype also gave us crucial insights into shortcomings of the current search interface. Even though more extensive training would have reduced the number of issues reported, it was highlighted where and how the system lacks in order for inexperienced users to quickly grasp the finer points of the search system. Based on this feedback, a redesign effort has been undertaken (cf. D6.8, Section 6.3.3) and these improvements are being implemented as part of the Flow product roadmap for Q2 of 2021.

Furthermore, shortcomings on searching and indexing combined modalities of temporal metadata were observed. The availability of more AME metadata from various modalities (speech, faces, OCR, etc.) exposed shortcomings to the indexing system that underlies the Flow system. Combining a single temporal modality (e.g., ASR transcripts) with metadata associated with entire clips works perfectly, searching starts to break down when users attempt to combine multiple temporal modalities (e.g., word x is spoken at the same time as person Y is visible in the image). These limitations can be worked around but still provide a significant hurdle for optimum content retrieval. We devised a strategy to rework our search indexing system to resolve these issues, and discussed it in Section 6.3.2 of D6.8. The actual implementation of this strategy however is still future work, but will be crucial when we want to bring the use of multi-modal querying and content retrieval into production use. Another approach to help tackle this problem is by introducing even richer multi-modal content descriptions (i.e., with identified people, actions, places, and summaries of spoken text) in a single metadata stratum which can more easily be indexed and searched through.

Despite these shortcomings, we did demonstrate how the search system could be elegantly adapted for reliable cross-lingual content retrieval, as discussed in D4.4. This piloted cross-lingual retrieval functionality will hence be incorporated into future improvements of the search system infrastructure.

10.2.3 Overall conclusions concerning the evaluation of "Intra- and interlingual subtitling"

Concerning subtitling, the process study gave valuable data on what happens in the process, and the process metrics indicate that post-editing ASR or MT can in fact increase productivity in intra- and interlingual subtitling. This despite the fact that the test panels did not clearly grade the process as being more efficient or enjoyable, or as always having great trust in the accuracy of the provided auto-generated source materials.

For intralingual subtitling, ASR with post-editing shows promise as a workflow, with most participants indicating they would be interested in using it further. This is exemplified by the fact that, thanks to the subtitling evaluations and the subtitling PoC (cf. also D7.4), YLE has decided to adopt ASR as an automation step into their intralingual subtitling process on a daily basis.

A clear observation is that even as MeMAD has pushed the state-of-the-art, it remains hard to beat the performance of professional subtitlers starting from scratch, even if they have a good ASR input and the spotting process automated. We have shown that the most glaring spotting and timing issues have been resolved during the project, but work still remains to improve the corrective burden left with the subtitler. This also manifests itself when considering current ASR transcription results. ASR implementations produce literal results and are often too correct for subtitling purposes as they include any hesitations and repetitions in the original speech. Leaving these elements out of subtitles (as is common practice), would already reduce the amount of post-editing significantly.

On the other hand, given the test panel's positive attitude towards many aspects of the automation process, incl. the quality of ASR, timing, etc., auto-subtitling is becoming a viable tool for non-professional subtitlers to quickly produce subtitles, either resulting in a somewhat lower standards result, or with a high standard but produced within a timeframe that is much shorter than it would take them to begin subtitling from scratch.

For interlingual subtitle post-editing by professional subtitlers, response from the participants was more mixed, although some interest was indicated toward MT and post-editing at least for some content types with further improvements in the output. The use of pre-existing intralingual subtitles as the source text for MT appears a feasible approach, although further improvements in quality and usability are needed. We are not yet at the point where clear productivity wins are to be obtained with the final implementations delivered in the project, even though good results were obtained for some language pairs between which machine translation was employed. Translations to the native language of the subtitler in particular suffered in the UEQ assessments with little positive feedback concerning the entire ordeal, while non-native translations were rated much higher in terms of quality and experience.

Starting interlingual subtitling from existing intralingual subtitles is clearly more straightforward than automating the process from scratch. Reusing correct choices of wording and grammar from the source language combined with timing from pre-existing subtitles eases the post-editing process significantly. The approach for combining ASR and MT into a pipeline that delivers interlingual subtitles from a raw audio signal suffers from second-order errors with errors that propagate from the ASR process and are then amplified by the MT which often turns a bad transcription into a worse translation, as also noted by the consumer reception test panels. Despite a potential comical effect, this situation hinders the adoption of this kind of subtitle generation on a large scale at the moment. Improvements are still required to bring this technology into production, unless perhaps clear disclaimers are provided that stress the fact that subtitles are auto-generated. We look forward to further research on process-aware ASR and MT implementations that better understand eithers outputs and requirements such that better translations could be obtained out-of-the-box in the future.

On the other hand, when we consider the feedback from the consumer reception trials, we do see encouraging results. The quality of the subtitles delivered by a fully automated chain are clearly not yet high enough, but they do provide crucial insights into foreign language content that would have otherwise remained inaccessible to a large audience. Interpreting these subtitles does require additional effort, but they often manage to bring across the correct message, without the need for any human curation intervention.

We note also that the cases investigated were the most extreme ones. Exploiting existing intra-lingual subtitles - that are often authored as part of satisfying accessibility legislation - can just a well serve as the source for large-scale unsupervised machine-translated subtitle generation. As the results from such a pipeline were judged much more positively by the professional subtitler panel, we can expect a similar better assessment from audiences. As such, this approach could be the first step in implementing these systems with the aim of enlarging audiences for content previously inaccessible because of the language barrier.
As a final remark, we note that the feasibility of automated interlingual subtitling will be highly language-dependent too, at least for the time being. ASR, MT and natural language processing (NLP) tools are on a higher readiness level for major languages than for the languages (consciously) evaluated in this project. Subtitle translations between well-serviced language pairs should produce better outcomes until the gap with lower-resourced languages is reduced, which we have proven feasible in the project (cf. subsection 10.1.1 and 10.1.4).

Impact on future development of the MeMAD prototype platform

With regard to the impact the subtitling evaluations have on future developments of the MEMAD platform and the Limecraft Flow commercial platform from which it is derived, we can already identify the following tracks of work:

- We keep improving the spotting and segmentation algorithms for the automated subtitle generation. While much improved over the course of the project, changes can be made to better support per-language preferences of spotting and increasing robustness for spotting across translations. The latter could mean that words are reshuffled during the MT stage, which could lead to unexpected spotting effects. This will be addressed wherever possible;

- Requests have also been raised about supporting organization-specific styling rules that go beyond regular subtitle spotting rules already implemented. Rules such as how currencies are written and abbreviations are used will have to find their place alongside the existing algorithms that employ natural language processing (NLP) to provide reasonable-looking subtitles (cf. D6.5);
- We will investigate how the ASR+MT combination can be tuned to reduce second-order errors that propagate between the two processes. Another research avenue will be to explore how limited text rewriting and condensation can be adopted to better match the nature of subtitling, as opposed to the literal transcription offered by ASR technology today. Naturally, assuring that such functionality rewrites source text with respect to the meaning of the source input is crucial;
- We will work on improving the subtitling GUI offered by the platform to facilitate more convenient retiming and splitting/merging of subtitles, features that are now more convenient in use in other dedicated subtitling software.

10.2.4 Overall conclusions concerning the evaluation of "Auto-generation and correction of content descriptions"

The evaluation of the content description editor prototype has shown that we have a solid foundation to work towards a usable next-generation production tool. Especially if the intended improvements (cf. Section 9.3.1) to the prototype are implemented in the near future.

Even if the prototype tool itself obtains a high level of ease of use and allows curators to fluently produce accurate content descriptions, challenges remain to improve the services that feed the tool with input from which the human curators start their work. If we can make the visual captioning perform better, with regards to even more content modalities and the editorial context in which it is employed, the amount of correction required will be reduced further. Examples include taking into account the expected story grammar (as suggested in D5.3) and incorporating domain-specific expert know-how into the captioning process. This is work is ongoing after the PoC with AP (cf. D7.4), in which we strive to deliver descriptions of shot lists to story editors that require little to no correction before publication. Similarly, if we can improve the accuracy of content segmentation (cf. D6.8 and D3.3), it can serve as an important first pass of dividing content into segments that need a human-verified description. If automated content description can improve significantly, audio description, one of the use cases originally envisioned for MeMAD but deemed unrealistic in this project, could once more be revisited in the future.
Better content understanding in captioning will also lead to more robust implementations of the auto-generation of stories functionality that we laid the foundation for in our prototype (cf. Section 6.5 of D6.8). This in turn can lead to more elaborate pilots with producers such as KRO-NCRV that are looking for these solutions right now (cf. D7.4).

Impact on future development of the MeMAD prototype platform

Work remains to determine how the workflow built from the content description application and the captioning technology that supports it can be further enhanced such that commercial adoption becomes feasible. We have a good grasp of the technical challenges involved, but the business case still needs to be clarified. We also need to determine to what extent our prototype tool needs to be adapted for customizations to better fit in with customer's archives and content annotation workflows and guidelines. We will continue our conversions with the organizations whose people participated in the content description editor evaluation to assess the potential for follow-up pilot projects and possible integrations.

Research-wise, future work will also need to further close the gap between the content search and retrieval functionality (i.e., Epic 6.3) on one side and (automated) content description on the other. These are closely intertwined as the result from the latter will be used as input by the former. Questions including where the balance lies between how much human curation is still needed given the availability of a blanket of AME metadata remain to be determined in the future.

11 Conclusions and dissemination activities

This deliverable concludes the work of MeMAD's WP6 and describes the last set of evaluation rounds concerning the prototype MeMAD platform as described by the complement of D6.2, D6.5 and D6.8. This deliverable reported on the user feedback and observations gathered by the consortium during the user panel evaluations that took place during the final fifteen months of the project.

A strong collaboration between project consortium partners Limecraft, YLE, UH and Surrey was setup to successfully execute many evaluation studies across many different use cases implemented in the MeMAD project. In particular, we completed 6 extensive evaluations in addition to the first round of 4 evaluations in the previous round. In total, over 100 participants, both professionals from the media production industry – from within the consortium, from our External Collaborators Group member organizations and from other interested parties – and consumers from a variety of backgrounds participated in the evaluation sessions organized by WP6.

These evaluations presented us with clear conclusions on the feasibility of adopting the functionality implemented as a prototype during the later dissemination and exploitation of the MeMAD results. As the same time, they presented us with suggestions for improvements and with insights into potential productivity gains for those workflows that have reached a level of maturity in which such quantitative measures could be employed. In all other cases we managed to gather reliable qualitative feedback on which aspects of the presented technologies, workflows and user interfaces worked or did not work in helping people perform their media production duties using the MeMAD integrated prototype platform.

The observations made in the evaluations of WP6, together with our findings from the proofs-of-concept discussed in D7.4 will now be utilized alongside the business and exploitation plan (cf. D7.2) to find the proper commercial application of the technologies developed in MeMAD. This report will serve as input to the development roadmap involved in this effort, and should also indicate where further fundamental research is required to bring promising technologies into the everyday workflow of many media production professionals.

11.1 Dissemination activities

The following dissemination activities have taken place for the final MeMAD prototype and the evaluations described in this deliverable (i.e., for the period M25 – M39 of the project).

- **Presentation:** 10/06/2020: EBU MDN 2020: EBU Metadata Developer Network Workshop, (online, hosted from Geneva, Switzerland): *Metadata Processing in the H2020 MeMAD platform*, by Dieter Van Rijsselbergen.
- **Publication**: Lauri Saarikoski, Dieter Van Rijsselbergen, Maija Hirvonen , Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen: *MeMAD project: End User Feedback on AI int the Media Production Workflows*, International Broadcasting Convention (IBC) 2020 Conference, proceedings are available online at: https://www.ibc.org/technical-papers/memad-project-end-user-feedback-on-ai-in-the-media-production-workflows/6764.article .
- **Presentation:** 27/10/2020: Joint IASA - FIAT/IFTA conference 2020: EBU (online): "*Where to apply AI? Measuring the value of automated metadata and machine translations in professional media archive use*" by Lauri Saarikoski and Dieter Van Rijsselbergen.
- **Presentation and demonstration**: 3/12/2021: European Language Grid Meta-forum 2020 (online): "Demonstration of the MeMAD project demonstrator prototype platform and findings from user evaluations" and demonstration of the MeMAD prototype between presentation sessions, by Dieter Van Rijsselbergen.
- **Presentation:** 27/01/2021: EBU PTS 2021: EBU Production Technology Seminar (online and hosted from Geneva, Switzerland): "*AI in Media Production – Findings from MeMAD*" by Dieter Van Rijsselbergen.
- **Presentation:** 02/02/2021:MeMAD Webinar series: "*Industrialising Media Production – A Producer's Perspective*" by Maarten Verwaest, Michael Strombom and Dieter Van Rijsselbergen, webinar recording available online at: https://memad.eu/webinars/production/.
- **Presentation:** 21-23/06/2021: At the 2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021): "*AI in Media Production - What works now, and which challenges still need solving*", keynote talk by Dieter Van Rijsselbergen.

12 Bibliography

[1] W. Tan, D. Liu, R. R. Bishu, A. Muralidhar and J. Meyer, "Design improvements through user testing," in *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting, 1181-1185.*, 2001.

[2] B. Laugwitz, T. Held and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire.," in *In HCI and Usability for Education and Work. USAB 2008, edited by Andreas Holzinger. Lecture Notes in Computer Science.*, Berlin, Heidelberg, Springer, 2008, p. 5298:63–76.

[3] B. Matthews and L. Ross, Research Methods: A Practical Guide for the Social Sciences, Edinburgh: Pearson Education Ltd., 2010.

[4] S. Wilkinson, "Analysing Interaction in Focus Groups," in *Talk and Interaction in Social Research Methods*, P. Drew, G. Raymond and D. Weinberg, Eds., London/Thousand Oaks/New Delhi, SAGE Publications, 2006, pp. 50-62.

[5] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba and A. Oliva, "SUN Database: Exploring a Large Collection of Scene Categories," *International Journal of Computer Vision,* vol. 119, no. 2016, pp. 3-22, 2014.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and a. E. Herbst, "Moses: open source toolkit for statistical machine translation.," in *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*, Association for Computational Linguistics, USA., 2007.

[7] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins and A. Birch, "Marian: Fast Neural Machine Translation in C++," in *In Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018.

[8] K. A. Ericsson and H. A. Simon, Protocol analysis: Verbal reports as data, The MIT Press, 1993.

[9] M. Leijten and L. Van Waes, "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes," *Written Communication,* vol. 30, no. 3, p. 358–392, 2013.

[10] J. Rubin and D. Chisnell, Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, Indianapolis, IN: Wiley, 2nd edn., 2008.

[11] M. van den Haak, M. De Jong and P. J. Schellens, "Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue," *Behaviour & Information Technology,* vol. 22, no. 5, pp. 339-351, 2003.

[12] M. van Someren, Y. Barnard and J. Sandberg, The think aloud method: A practical guide to modelling cognitive processes, Academic Press, 1994.

Appendix A Epic 6.11 – Intralingual subtitling: Proof-of-concept feedback form, evaluation post-task questionnaire and post-evaluation interview questions

Intralingual subtitling - PoC feedback form

Programme
Enter the name and media ID of the programme you worked on. Please note if you used ASR for only a part of the programme, eg. “A-studio, the first 20 minutes, [MEDIA ID]”

Programme genre
Eg. current events, drama, lifestyle

How much time did you spend on the task?
Estimate the time you spent on the task, starting from when you received the ASR subtitling template and ending when the subtitles were complete.

Post-editing was…
[7-point Likert scales]
Easy - Difficult
Unpleasant - Pleasant
Relaxed - Stressful
Laborious - Effortless
Fast - Slow
Efficient - Inefficient
Boring - Exciting
Fun - Tedious
Complicated - Simple
Enjoyable - Annoying
Limiting - Creative
Motivating - Demotivating
Impractical - Practical

The pre-made subtitling template
[7-point Likert scales]
The automatic timecoding was poor - The automatic timecoding was good
Fixing the timecoding took a lot of effort - Fixing the timecoding was easy
The automatic segmentation was poor - The automatic segmentation was good
Fixing the segmentation took a lot of effort - Fixing the segmentation was easy
The quality of the ASR was poor - The quality of the ASR was good
Correcting the recognition errors took a lot of effort - Correcting the recognition errors was easy

Describe in your own words what the post-editing experience was like

What kind of errors were there in the ASR results?
Were there any recurring errors? What kind of recognition errors were the most trouble in post-editing? Give examples of errors and their corrections.

What kind of errors were there in the timecoding and segmentation?
Were there any recurring errors? What kind of segmentation and timecoding errors were the most trouble in post-editing? Give examples.

Upload the completed text file here in .xml format

Intralingual subtitling - Post-task questionnaire

Task identifier
Enter the name of the task you just completed

Time spent on task
How much time did you spend on the task, in minutes?

Work phases
Check all the work phases you completed, outside of the post-editing

- ❏ Watching the video without the ASR template before starting the post-editing
- ❏ Watching the video with the ASR template before starting the post-editing
- ❏ Watching the video with the post-edited subtitles
- ❏ Other…

Retrospection

Tell us briefly about your observations concerning the ASR template and post-editing of this clip. The purpose of these questions is to report what you remember thinking while you were working on the subtitling task. Watch the clip once more to refresh your memory, but do not change anything at this stage. You are not expected to redo the translation, but to tell us what you remember thinking while you edited the machine translation. Try to tell us which moments, words, expressions etc. your observations were about, in as much detail as possible.

What did you think about the quality of the automatic speech recognition? (Try to ignore segmentation and timecoding at this time.) What kind of problems were there in the ASR? What was good about it?

What did you think about the quality of the automatic segmentation/timecoding? (Try to ignore the translation itself at this time.) What kind of problems were there in the segmentation/timecoding? What was good about it?

Post-editing was…
[7-point Likert scales]
Easy - Difficult
Unpleasant - Pleasant
Relaxed - Stressful
Laborious - Effortless
Fast - Slow
Efficient - Inefficient
Boring - Exciting
Fun - Tedious
Complicated - Simple
Enjoyable - Annoying
Limiting - Creative
Motivating - Demotivating
Impractical - Practical

The pre-made subtitling template
[7-point Likert scales]
The automatic timecoding was poor - The automatic timecoding was good
Fixing the timecoding took a lot of effort - Fixing the timecoding was easy
The automatic segmentation was poor - The automatic segmentation was good
Fixing the segmentation took a lot of effort - Fixing the segmentation was easy

Were there differences between these post-editing tasks or ASR templates and the ones you saw in the 2019 evaluations? What were they?

What was the video like as a task? Was there anything particularly difficult or laborious in the task that did not have to do with the ASR template?

Upload the final subtitle file here

Intralingual subtitling - semi-structured interview

What’s your overall impression about this ASR experiment?

(If they do not explain:) Why is that?

Did you notice ASR working better in some tasks compared to others?

Did you notice any difference to the previous round of evaluations (in 2019)?

Would you use this Flow or a similar platform as a tool for subtitling?

How should ASR be developed so that it would be a better tool?

(Additional question, to be used if the participant has answered only briefly:) Is there anything else you would like to say about this experiment?

Appendix B Epic 6.11 – Interlingual subtitling: Evaluation post-task questionnaire and post-evaluation interview questions

Interlingual subtitling - post-task questionnaire

Task identifier
Enter the name of the task you just completed, e.g. FEA-1

Time spent on task
How much time did you spend on the task, in minutes?

Work phases
Check all the work phases you completed, outside of the post-editing

- ❏ Watching the video without the machine translation before starting the post-editing
- ❏ Watching the video with the machine translation before starting the post-editing
- ❏ Watching the video with the post-edited subtitles
- ❏ Other…

Retrospection

Tell us briefly about your observations concerning the machine translation and post-editing of this clip. The purpose of these questions is to report what you remember thinking while you were working on the subtitling task. Watch the clip once more to refresh your memory, but do not change anything at this stage. You are not expected to redo the translation, but to tell us what you remember thinking while you edited the machine translation. Try to tell us which moments, words, expressions etc. your observations were about, in as much detail as possible.

What did you think about the quality of the automatic translation? (Try to ignore segmentation and timecoding at this time.) What kind of problems were there in the translation? What was good about it?
If you wish, you may refer to the text file with the original machine translation to refresh your memory

What did you think about the quality of the automatic segmentation/timecoding? (Try to ignore the translation itself at this time.) What kind of problems were there in the segmentation/timecoding? What was good about it?
If you wish, you may refer to the text file with the original machine translation to refresh your memory

[New section]

Post-editing was…
[7-point Likert scales]
Easy - Difficult
Unpleasant - Pleasant
Relaxed - Stressful
Laborious - Effortless

Fast - Slow
Efficient - Inefficient
Boring - Exciting
Fun - Tedious
Complicated - Simple
Enjoyable - Annoying
Limiting - Creative
Motivating - Demotivating
Impractical - Practical

Machine translated subtitles
Respond according to what feels right; do not stop to think about the questions too much, and follow your intuition
[7-point Likert scales]
The pre-made timecoding was poor - The pre-made timecoding was good
Fixing the timecoding took a lot of effort - Fixing the timecoding was easy
The pre-made segmentation was poor - The pre-made segmentation was good
Fixing the segmentation took a lot of effort - Fixing the segmentation was easy

Any other comments

Upload the completed subtitling file here, in .srt format

Interlingual subtitling - semi-structured interview

What’s your overall impression about these post-editing tasks?

(If they do not explain:) Why is that?

Did you notice any differences between the machine translations?

Did you notice any difference to the previous round of evaluations (in 2019)?

Could you imagine using machine translation as a tool for subtitling?

Do you think some other form of metadata would have been useful for you, such as speech recognition transcripts without machine translation, or facial recognition of the people in the video?

How should machine translation be improved for it to be a better tool?

(Additional question, to be used if the participant has answered only briefly:) Is there anything else you would like to say about this experiment?

Appendix C Epic 6.11 – Interlingual subtitling – consumer reception: Focus group script 1 (16/06/2020), focus group script 2 (27/10/2020), questionnaire “on viewing video clips with machine-translated subtitles” and post-focus group feedback questionnaire.

Interlingual subtitling - consumer reception - Focus group 1 - English

16 June 2020

Introduction and instructions

Welcome to our group discussion, and thank you for participating. My name is Tiina Tuominen, I'm a researcher at Yle, and I will be moderating this discussion. We also have a second researcher present, her name is Maarit Koponen and she is from the University of Helsinki. She doesn't intend to participate in the conversation, but she will be observing it and helping me with technical or other issues if necessary.

Before we start, let's check that the technology is working properly. Can everyone see and hear me? *Chat window: If you can't hear my voice, please post a message to let me know.* Could everyone please take turns and say Hello or something so that I can check that I hear everyone and you hear each other. If at any point you lose either the sound or the image, alert me to it via the chat window. You can also use the chat window to request a speaking turn if it is otherwise difficult to get a word in. You can just post number 1 and I'll know what it means. If my internet connection drops off, just wait a moment and I'll try to be back as soon as I can. If your connection disappears, please do the same and click on the meeting link again. I will be video recording this session, and I'm starting the recording now.

You have received preliminary information about this session, but just to recap, we are conducting research on the use of technological tools in translating and subtitling tv programmes, making the process faster and more automatic. We are now trying to find out what viewers think of subtitles that have been created in a new way. These subtitles would not necessarily replace tv subtitles, they would be for online videos and similar instances where we want to get the material out to the public quickly.

What we are going to do in this session is that after a warm-up question, I will show you a short video clip. It is about five minutes long, the language is Finnish, and it is subtitled in English. It is a current affairs programme. After we have watched the video, I'll ask a few questions about it. The questions will be about the contents of the video (nothing complicated, though), about how you felt about watching the video, how much you understood and so on.

I want to emphasise that we are looking for input on how watchable the video is from your point of view. There are absolutely no wrong answers, and any and all comments are welcome. If you have trouble understanding the video, it is probably the video's fault, not yours. And because we are calling this a discussion, the idea is that I will ask my questions as conversation starters, and you are then free to discuss each topic amongst yourselves, expand on each others' answers, agree, disagree, anything. So please feel free to keep talking, and don't wait for my permission. It's okay if the conversation starts having a life of its own, I will bring it back to the topic if and when necessary. And remember that you can use the chat window to request a turn to speak if you need to.

The research information sheet described how we are using the focus group data and your personal information. If you are now prepared to start the actual focus group discussion, that is an indication that you accept all that. Of course, you are free to withdraw from the study afterwards if you decide that you want to. Any questions at this point?

Discussion

If you don't have any more questions, we can start. Before the video, I'll ask a warm-up question to get us into the topic.

Do you watch a lot of Finnish-language programmes that have been subtitled into English? Do you find them easy to follow?

You can take turns answering, and if you'd like, you can comment on each other's responses.

10 min

I'll start the video now. There is about a minute without any speaking at the start, and then a few seconds before the subtitles start after the narration starts, but I promise it will start eventually. If you have any problems with the sound or the image, please let me know in the chat so I can press pause. It may improve the quality of the video if you switch off your webcams for the duration of the video by clicking on the video camera at the bottom of the screen. Although it's not really a problem if you don't hear the sound well, because you are not expected to understand it anyway.

https://youtu.be/ONhwv0Awa5k

20 min

Now that we've seen the video, let's start with a very general question. **How was it to watch this clip? What are your immediate thoughts?**
Did it feel easy or difficult to follow the clip? (Was it easy or difficult to read the subtitles?)

What do you think of the subtitles? How would you assess them on a scale of 1-10?
Would these subtitles be good enough for you to watch these types of videos online?

30 min

Then, let's talk about the contents of the video. **Did you understand what the clip was about?**
Could you describe briefly what was in the video? Does everyone agree?

40 min

Then I have a few questions about your experience of watching the clip.
Was the pace of the clip suitable, or too fast or too slow?
(Were things expressed in a clear or unclear way?)
(Do you feel that you now know more about the topic of the clip?)
Do you feel you had to invest a lot of mental effort in gaining new information out of the clip?

50 min

As you may have guessed, this was a machine translation. **What do you think about that? If your options are either this type of translation immediately or a human translation or a journalistic piece in English the following day, which one would you choose? Imagine that this was about something current in the news.**

How would you like to see the subtitles improved?

Have you changed your mind about the number assessment or any of your other comments? E.g. Do you think the subtitles are good enough to watch these types of clips online?

65 min

We are almost out of time. **Do you have any other comments, questions, observations?**

If there is nothing else, we can finish the discussion, but first I have a couple of announcements. First, the research project is called MeMAD. If you'd like to find out more about it, I'll post a link to our website in the chat.

https://memad.eu/

I'm also posting a link to a small, anonymous questionnaire, in case you want to add anything to your answers or give us feedback on the study. It is voluntary, but it's there for you if you have something on your mind or think of something afterwards that you would have liked to say. The questionnaire will be available until 25 June, 6 pm.

https://forms.gle/sbEn9CYpgV8etuf16

You might want to copy paste the links to a text document so that you can have a look at them afterwards, or you can just click on the links and then bookmark them or something. You can't access the chat after we finish.

Thank you very much to all of you for participating. If you have any questions afterwards, in addition to the questionnaire, you are very welcome to contact me, Maarit or Kaisa directly.

Our contact information is in the research information leaflet. Thank you again, you can click off now.

Interlingual subtitling - consumer reception - Focus group 2 - English

27 October 2020

Introduction and instructions

Welcome to our group discussion, and thank you for participating. My name is Tiina Tuominen, I'm a researcher at Yle, and I will be moderating this discussion. We also have a second researcher present, her name is Maarit Koponen and she is from the University of Helsinki. She doesn't intend to participate in the conversation, but she will be observing it and helping me with technical or other issues if necessary.

Before we start, let's check that the technology is working properly. Can everyone see and hear me? *Chat window: If you can't hear my voice, please post a message to let me know.* Could everyone please take turns and say Hello or something so that I can check that I hear everyone and you hear each other. If at any point you lose either the sound or the image, alert me to it via the chat window. Keep your sound on mute when you are not talking. When you want to talk during the discussion, unmute yourself, and I'll see that and know to give you a turn. If my internet connection drops off, just wait a moment and I'll try to be back as soon as I can. If your connection disappears, please do the same and click on the meeting link again.

You have received preliminary information about this session, but just to recap, we are conducting research on the use of technological tools in translating and subtitling audiovisual materials, making the process faster and more automatic. In practice this means that we are using machine translation. We are now trying to find out what viewers think of these automatic subtitles. These subtitles would not replace tv subtitles, they would be for online videos and similar instances where we want to get the material out to the public quickly.

What we are going to do in this session is that after a warm-up question, I will show you two short video clips, about three minutes each. The language is Finnish, and they are subtitled in English. The first one is a news story and the second a clip from a current affairs programme. After we have watched the first video, I'll ask a few questions about it, then we'll watch the second video and talk about it, and finally I'll have some questions about the experience in general. The questions will be about the contents of the videos (nothing complicated, though), about how you felt about watching the video, how much you understood and so on.

I want to emphasise that we are looking for input on how watchable the videos are from your point of view. There are absolutely no wrong answers, and any and all comments are welcome. If you have trouble understanding the videos, it is probably the videos' fault, not yours. And because we are calling this a discussion, the idea is that I will ask my questions as conversation starters, and you are then free to discuss each topic amongst yourselves, expand on each others' answers, agree, disagree, anything. So please feel free to keep talking, and don’t wait for my permission. It’s okay if the conversation starts having a life of its own, I will bring it back to the topic if and when necessary.

The research information sheet described how we are using the focus group data and your personal information. If you are now prepared to start the actual focus group discussion, that is an indication that you accept all that. Of course, you are free to withdraw from the study afterwards if you decide that you want to. Any questions at this point? I will be video recording this session, and I’m starting the recording now.

Discussion

Before the video, I’ll ask a warm-up question to get us into the topic.

Do you watch a lot of Finnish-language programmes that have been subtitled into English? Do you find them easy to follow?

You can take turns answering, and if you’d like, you can comment on each other’s responses.

10 min

I’ll start the video now. The speaking and the subtitles start pretty immediately. If you have any problems with the sound or the image, please let me know in the chat so I can press pause. Please switch off your webcams for the duration of the video by clicking on the video camera at the bottom of the screen, it may help the video run more smoothly. Obviously it’s not really a problem if you don’t hear the sound well, because you are not expected to understand it anyway.

https://youtu.be/ql-yb6lRH9o

15 min

Now that we’ve seen the video, let’s start with a very general question. **What did you think of the video? How was it to watch this clip?**

(Did it feel easy or difficult to follow the clip?)

What do you think of the subtitles?
Would these subtitles be good enough for you to watch these types of videos online?

Did you understand what the clip was about? Could you describe briefly what it was about? Does everyone agree?

30 min

Any other comments about this video? If not, let's watch the second one. Please switch off your webcams again.

https://youtu.be/apzvQP88SyE

35 min

Now that we've seen the video, let's start with a very general question. **What did you think of the video? How was it to watch this clip?**

(Did it feel easy or difficult to follow the clip?)

What do you think of the subtitles?
Would these subtitles be good enough for you to watch these types of videos online?

How was this video compared to the previous one?

Did you understand what the clip was about? Could you describe briefly what it was about? Does everyone agree?

50 min

Let's talk about both clips in general then. So these were both machine translations, and the second one was slightly polished by a human (that may not be very noticeable, but e.g. some of the names were fixed). **If your options are either the first video immediately, the second one with a slight delay, or a human translation or a journalistic piece in English in a couple of days, which one would you choose? Imagine that this was about something current in the news.**

How would you like to see the subtitles improved?

Can you imagine a topic or situation where you could watch something with subtitles like these?

(Have you changed your mind about anything during the discussion? E.g. Do you think the subtitles are good enough to watch these types of clips online?)

60 min

We are almost out of time. **Do you have any other comments, questions, observations?**

If there is nothing else, we can finish the discussion, but first I have a couple of announcements. First, the research project is called MeMAD. If you'd like to find out more about it, I'll post a link to our website in the chat.

https://memad.eu/

I’m also posting a link to a small, anonymous questionnaire, in case you want to add anything to your answers or give us feedback on the study. It is completely voluntary, but it’s there for you if you have something on your mind or think of something afterwards that you would have liked to say. The questionnaire will be available until Thursday 5 November, 6 pm. Here is the link:

https://forms.gle/uztCTWhdHi8HSBHQA

You might want to copy paste the links to a text document so that you can have a look at them afterwards, or you can just click on the links and bookmark them or something. You can’t access the chat after we finish.

Thank you very much to all of you for participating. If you have any questions afterwards, in addition to the questionnaire, you are very welcome to contact me, Maarit or Kaisa directly. Our contact information is in the research information leaflet. Thank you again, you can leave the session now.

Survey on viewing video clips with machine-translated subtitles

Thank you for taking the time to respond to our survey! The responses will be used for research purposes only. No personal information will be collected in the questionnaire. Data are processed and stored in a protected location in an anonymised format so that an individual participant cannot be identified. The anonymous data may be shared with other parties in the research project, and researchers participating in the project may continue to use the anonymised data in research and research publications after the end of the project. Information collected through this form is stored on servers administered by Google, which may be located outside the European Union/European Economic Area. The information is protected by a user account and password. Information will be used only by researchers in the MeMAD project. The data controller is Yleisradio, represented here by Tiina Tuominen (tiina.tuominen@yle.fi).

Participation in this research is voluntary. You may stop responding to the questionnaire at any point, in which case none of your answers will be saved or used in the study. When you press Send at the end of the questionnaire, your answers will be added to the research data, and it will no longer be possible to remove them. In other words, by pressing Send you are giving permission for your answers to be used in the study.

If you have any questions about the study, you can contact the researchers Tiina Tuominen, Kaisa Vitikainen or Maarit Koponen at the addresses given below.

Many thanks for contributing to our study!

Dr Tiina Tuominen
Developer of Translation and Subtitling
P.O. Box 15, FI-00024 Yleisradio
tiina.tuominen@yle.fi

Kaisa Vitikainen
Yle Translations
P.O. Box 62, FI-00024 Yleisradio
kaisa.vitikainen@yle.fi

Dr Maarit Koponen
Department of Digital Humanities
University of Helsinki
maarit.koponen@helsinki.fi
*Required

1. I have received sufficient information about the research and handling of research data, and I am willing to take part in the study. *

   Mark only one oval.

   ○ Yes

2. I am at least 18 years old. *

   Mark only one oval.

   ○ Yes

Information about you

In this section, we would like you to share some details about yourself that are relevant to the study.

3. What is your age? *

   Mark only one oval.

   ○ 18–29
   ○ 30–44
   ○ 45–59
   ○ Over 59
   ○ I prefer not to say

4. What is the highest level of education you have completed? *

Mark only one oval.

○ I have not completed any level of education

○ Primary education

○ Secondary education

○ Vocational qualification

○ Some studies at a university or other institute of higher education, but no degree

○ Undergraduate degree (e.g. BA, BSc)

○ Postgraduate degree (e.g. MA, MSc)

○ Doctoral degree or higher

○ Other: ______________________

5. How well do you understand spoken Finnish? *

Mark only one oval.

○ Not at all

○ I understand some words and phrases in simple spoken Finnish

○ I understand the main points in simple spoken Finnish

○ I understand the majority of simple spoken Finnish

○ I understand even complex spoken Finnish quite well

○ I usually have no difficulty understanding any kind of spoken Finnish

6. Do you live or have you ever lived in Finland? *

Mark only one oval.

○ I currently live in Finland and have lived in Finland for less than a year
○ I currently live in Finland and have lived in Finland for 1–5 years
○ I currently live in Finland and have lived in Finland for 6–10 years
○ I currently live in Finland and have lived in Finland for more than 10 years
○ I do not live in Finland, but I have previously lived in Finland for less than a year
○ I do not live in Finland, but I have previously lived in Finland for 1–5 years
○ I do not live in Finland, but I have previously lived in Finland for 6–10 years
○ I do not live in Finland, but I have previously lived in Finland for more than 10 years
○ I have never lived in Finland

7. How often do you watch foreign-language programmes with English subtitles? *

Mark only one oval.

○ Never
○ Rarely (a few times a year or less)
○ Occasionally (every month or nearly every month)
○ Often (every week)

Video 1

Please watch the video clip below and then answer the questions related to the video. If you want to, you can watch the video more than once, but you can answer the questions based on just a single viewing.

Video 1

http://youtube.com/watch?v=X8jMxmbkC7c

8. Did you understand what the topic of the video was and what was said about that topic?

Mark only one oval.

	1	2	3	4	5	
Not at all	○	○	○	○	○	Yes, completely

9. What is the Patria AMV? *

Mark only one oval.

○ A weapons system used by Rangers in the battlefield

○ An armoured military transport vehicle

○ The official name of the PASI vehicle

○ A mobile military intelligence system

○ I don't know

10. What was going on in the office in Vienna at the beginning of the clip? *

Mark only one oval.

○ Patria's Austrian agent was losing patience with Patria.

○ Patria's Slovenian clients were attempting to deceive the Austrian agent.

○ Patria was pressuring its Slovenian clients to pay for their purchases.

○ The Slovenian visitors were demanding money from Patria.

○ I don't know

11. What was the suspicious activity discovered in Austria? *

Mark only one oval.

○ Industrial espionage
○ Suspicious travel patterns by known criminals
○ Questionable international money transfers
○ Sales of potentially unlawful products
○ I don't know

Your views on the video and the subtitles

12. Was the video pleasant to watch? *

Mark only one oval.

	1	2	3	4	5	
Not at all pleasant	○	○	○	○	○	Very pleasant

13. How useful were the subtitles in helping you understand the clip? *

Mark only one oval.

	1	2	3	4	5	
Not at all useful	○	○	○	○	○	Very useful

14. Did the information in the subtitles seem accurate to you? *

Mark only one oval.

	1	2	3	4	5	
No, not at all accurate	○	○	○	○	○	Yes, completely accurate

15. In comparison to an average experience of viewing a subtitled programme, how well did you manage to read the subtitles all the way through? *

Mark only one oval.

	1	2	3	4	5	
Considerably worse than usual	○	○	○	○	○	Considerably better than usual

16. In comparison to an average experience of viewing a subtitled programme, how much mental effort did you have to invest to learn new information from the clip? *

Mark only one oval.

	1	2	3	4	5	
Considerably more than usual	○	○	○	○	○	Considerably less than usual

Video 2

Please watch the video clip below and then answer the questions related to the video. If you want to, you can watch the video more than once, but you can answer the questions based on just a single viewing.

Video 2

http://youtube.com/watch?v=sKuLuysW6YY

17. Did you understand what the topic of the video was and what was said about that topic?

Mark only one oval.

	1	2	3	4	5	
Not at all	○	○	○	○	○	Yes, completely

18. Why do towns in Eastern Finland want a faster train connection to Helsinki? *

Mark only one oval.

○ They hope it will bring in some much-needed investments to remote areas
○ They hope it means the state is willing to invest in public transport
○ They hope it will bring in larger numbers of tourists
○ They hope it will slow down population decline
○ I don't know

19. How would the new rail connection be financed? *

Mark only one oval.

○ A new state-owned company would pay for it
○ Local municipalities would invest in the project
○ The increasing passenger numbers would cover the costs
○ No one knows yet
○ I don't know

20. What is the current status of the project? *

Mark only one oval.

○ The plans have been made and construction will start soon
○ It is currently in the planning stage
○ No decisions have been made yet
○ The company is committed to completing the project by 2030
○ I don't know

Your views on the video and the subtitles

21. Was the video pleasant to watch? *

Mark only one oval.

	1	2	3	4	5	
Not at all pleasant	○	○	○	○	○	Very pleasant

22. How useful were the subtitles in helping you understand the clip? *

Mark only one oval.

	1	2	3	4	5	
Not at all usefull	○	○	○	○	○	Very useful

23. Did the information in the subtitles seem accurate to you? *

Mark only one oval.

	1	2	3	4	5	
No, not at all accurate	○	○	○	○	○	Yes, completely accurate

24. In comparison to an average experience of viewing a subtitled programme, how well did you manage to read the subtitles all the way through? *

Mark only one oval.

	1	2	3	4	5	
Considerably worse than usual	○	○	○	○	○	Considerably better than usual

25. In comparison to an average experience of viewing a subtitled programme, how much mental effort did you have to invest to learn new information from the clip? *

Mark only one oval.

	1	2	3	4	5	
Considerably more than usual	○	○	○	○	○	Considerably less than usual

Views on automatic subtitles

26. Did it feel more difficult to watch these automatically subtitled video clips than it is to watch regular subtitled foreign-language content? *

Mark only one oval.

	1	2	3	4	5	
Considerably more difficult	○	○	○	○	○	No more difficult

27. Under what circumstances could you imagine watching videos on platforms such as news websites or social media with subtitles like these? *

Mark only one oval.

○ Any time automatic subtitles are available for a foreign-language video that interests me.

○ Only if I cannot easily find similar videos or information on the same topic in a language I understand.

○ Only if the topic is exceptionally interesting or important and there is no information available about it in a language I understand.

○ Never

○ Other: ______________________

28. If you could choose either automatic subtitles that were available immediately or human-made subtitles that were available a couple of days later, which one would you choose? *

Mark only one oval.

○ I would definitely choose the automatic subtitles.

○ I would choose the automatic subtitles in most cases.

○ I could take either one.

○ I would choose the human-made subtitles in most cases.

○ I would definitely choose the human-made subtitles.

○ Other: ______________________

29. Can you think of some situations or some types of videos with which you could use automatic subtitles like the ones you saw in the video clips?

____

____

____

____

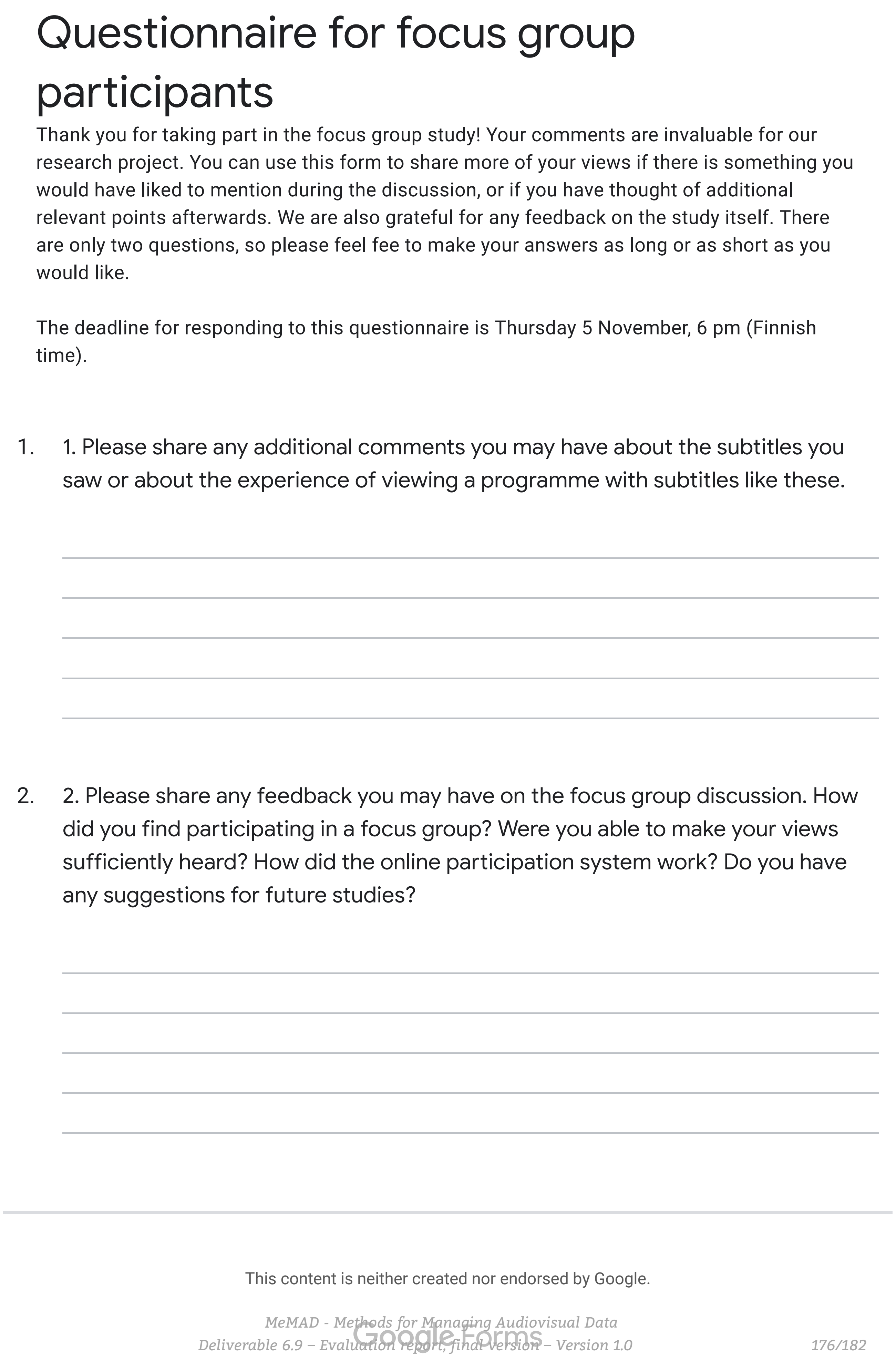
____

30. If you do not think these subtitles were good enough, what are some things that should be improved in order for you to use them?

Other comments and feedback

31. If you have any other comments about the subtitles you saw or about using automatic subtitles and machine translation, you can share them here.

32. This is where you can give feedback on this questionnaire.



Questionnaire for focus group participants

Thank you for taking part in the focus group study! Your comments are invaluable for our research project. You can use this form to share more of your views if there is something you would have liked to mention during the discussion, or if you have thought of additional relevant points afterwards. We are also grateful for any feedback on the study itself. There are only two questions, so please feel fee to make your answers as long or as short as you would like.

The deadline for responding to this questionnaire is Thursday 5 November, 6 pm (Finnish time).

1. 1. Please share any additional comments you may have about the subtitles you saw or about the experience of viewing a programme with subtitles like these.

2. 2. Please share any feedback you may have on the focus group discussion. How did you find participating in a focus group? Were you able to make your views sufficiently heard? How did the online participation system work? Do you have any suggestions for future studies?

Appendix D Epic 6.10 – Auto-generation and correction of content descriptions: UEQ Questionnaire.

MeMAD user experience study of creating semi-automated video descriptions (MeMAD UX)

Questionnaire (final version)

Participant ID:

1. What is your gender?

Female Male Other Prefer not to say

2. What is your age range?

18-19 20-29 30-39 40-49 50-59 60-69 70+

3. Which of these best describes your highest level of education?

Secondary school level Bachelors Masters PhD

Other - Please specify ______________

4. How many years of work experience do you have in content description?

1-5 years 6-10 years >10 years

5. In which country do you work?

____

6. How would you rate your expertise with content editing platforms?

No experience Novice Intermediate Advanced Expert

(A) User Experience

This part of the questionnaire consists of pairs of contrasting attributes that may apply to the environment. Please tick the box that best represents how you feel about the environment.

Example:

attractive		x						unattractive

This response would mean that you rate the application as much more attractive than unattractive.

Please decide **spontaneously**. Don't think too long about your decision to make sure that you convey your original impression. Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a box in **every line**.

It is your **personal opinion** that counts. Please remember: there is **no wrong or right answer**!

Please assess the <u>platform</u> by ticking one box per line.

	1	2	3	4	5	6	7		
annoying								enjoyable	7
not understandable								understandable	8
creative								dull	9
easy to learn								difficult to learn	10
valuable								inferior	11
boring								exciting	12
not interesting								interesting	13
unpredictable								predictable	14
fast								slow	15
inventive								conventional	16
obstructive								supportive	17
good								bad	18
complicated								easy	19
unlikable								pleasing	20
usual								leading edge	21
unpleasant								pleasant	22
secure								not secure	23
motivating								demotivating	24
meets expectations								does not meet expectations	25
inefficient								efficient	26
clear								confusing	27
impractical								practical	28
organized								cluttered	29
attractive								unattractive	30
friendly								unfriendly	31
conservative								innovative	32

(B) How was your experience of working in the environment?

	1	2	3	4	5	6	7		
My experience felt natural								My experience did **NOT** feel natural	33
I did **NOT** feel comfortable working in this environment								I felt comfortable working in this environment	34
I felt the environment had a positive impact on my performance								I felt the environment had a negative impact on my performance	35
I feel that the environment has **NOT** helped me to produce good descriptions								I feel that the editing environment has helped me to produce good descriptions	36

37. Further comments (optional)

38. Did you encounter any technical problems during your session?

No Yes

If Yes, please explain:

(C) Please specify your views about the functionality and features of the platform:

	1	2	3	4	5	6	7		
It was easy to access the application.								It was **NOT** easy to access the application.	39
The information provided in the application was **NOT** too technical.								The information provided in the application was too technical.	40
The 'adding a content description' feature was efficient/functional.								The 'adding a content description' feature was **NOT** efficient/functional.	41
The timeline was **NOT** useful for navigating the clip.								The timeline was useful for navigating the clip.	42
The timeline was useful for creating new content descriptions.								The timeline was **NOT** useful for creating new content descriptions.	43
The timeline zoom and pan was **NOT** easy to use.								The timeline zoom and pan was easy to use.	44
Selecting a time range using the SET IN and SET OUT buttons was intuitive.								Selecting a time range using the SET IN and SET OUT buttons was **NOT** intuitive.	45
The places, persons and other suggested tags were **NOT** useful.								The places, persons and other suggested tags were useful.	46
The suggestion in the 'spoken text' field' was useful.								The suggestion in the 'spoken text' field was **NOT** useful.	47
The sidebar with existing content descriptions was **NOT** clear.								The sidebar with existing content descriptions was clear.	48
The timeline lane showing places, persons, tags was useful.								The timeline lane showing places, persons, tags was **NOT** useful.	49
The 'Deep Caption' timeline lane was **NOT** useful.								The 'Deep Caption' timeline lane was useful.	50
The 'Shots' timeline lane was useful.								The 'Shots' timeline lane was **NOT** useful.	51
The 'Faces' timeline lane was **NOT** useful.								The 'Faces' timeline lane was useful.	52

The ‘OCR’ (text detected in screen) timeline lane was useful.								The ‘OCR’ (text detected in screen) timeline lane was **NOT** useful.	53
The ‘Transcript’ lane for the language spoken in the clip was **NOT** useful.								The ‘Transcript’ lane for the language spoken in the clip was useful.	54

55. Further comments (optional):

56. Would you recommend this environment to a colleague?

Yes No

THANK YOU!