



MeMAD Deliverable

D6.6 – Evaluation report, intermediate version

Version 1.1

Grant Agreement number	780069
Action Acronym	MeMAD
Action Title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding Scheme	H2020-ICT-2016-2017/H2020-ICT-2017-1
Version date of the Annex I against which the assessment will be made	8.5.2019
Start date of the project	1.1.2018
Due date of the deliverable	31.12.2019
Actual date of submission	13.03.2020
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 the MeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content

Authors in alphabetical order		
Name	Beneficiary	e-mail
Maija Hirvonen	University of Helsinki / Tampere University	maija.hirvonen@tuni.fi
Maarit Koponen	University of Helsinki	maarit.koponen@helsinki.fi
Lauri Saarikoski	Yle	lauri.saarikoski@yle.fi
Umut Sulubacak	University of Helsinki	umut.sulubacak@helsinki.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
Liisa Tiittula	University of Helsinki	liisa.tiittula@helsinki.fi

Document revisions			
Version	Date	Authors	Changes
0.1	03/02/2020	Maija Hirvonen, Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen	Initial version of evaluation result descriptions.
0.2	10/02/2020	Maarit Koponen, Kaisa Vitikainen Dieter Van Rijsselbergen	Additional sections on implemented functionality and impact of the evaluations, with completed task summary and test user setups.
0.3	17/02/2020	Maarit Koponen, Umut Sulubacak	Improved figures throughout the report.
0.4	22/02/2020	Dieter Van Rijsselbergen	First version for internal review.
0.5	04/03/2020	Dieter Van Rijsselbergen	All sections completed, ready for final internal review.
1.0	10/03/2020	Dieter Van Rijsselbergen	Final version with all reviews processed and appendices attached.
1.1	31/03/2021	Dieter Van Rijsselbergen	Improvement of included appendices after deliverable review.

Abstract

This is the second evaluation report of MeMAD prototype described in D6.5, the second of three, which 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. The main outcome of this evaluation round is a cautiously optimistic result in that the test subjects positively evaluate the potential of the implemented services, especially in subtitling, but that work remains to be done in improving a variety of hindrances to fully exploit the possible gain in usability and efficiency of the services developed in the project. Despite some user misgivings, we do note a clear efficiency improvement in the authoring process of interlingual subtitles. This evaluation round will be followed up by more evaluations of new and improved functionality in the final project year.

Contents

1	Introduction	6
2	Changes with regards to deliverable D6.3	7
3	Methodology for the MeMAD prototype platform development	8
4	Use cases functionality evaluated in the second MeMAD prototype platform.....	11
5	General evaluation setup and methodology.....	16
5.1	User Experience Questionnaire (in all cases)	17
5.2	Semi-structured interviews (in all cases).....	17
5.3	Think-aloud Protocols (for Epics 6.3 and 6.5)	17
5.4	Process data collection (for Epic 6.11)	18
5.5	Materials used in the evaluation.....	18
6	Evaluation of Epic 6.5: Editing assistance using multi-modal and multi-lingual metadata	19
6.1	Motivation	19
6.2	User test setup	19
6.2.1	Material used and testing environment.....	19
6.2.2	Participants	22
6.2.3	User data collection and tasks.....	22
6.3	Analysis of user data	22
6.3.1	User Experience Questionnaire	22
6.3.2	Feedback from interviews.....	23
7	Evaluation of Epic 6.2: Searching and browsing for ingested and archived content.....	26
7.1	Motivation	26
7.2	User test setup	27
7.2.1	Material used.....	27
7.2.2	Participants	31
7.2.3	User data collection and tasks.....	31
7.3	Analysis of user data	32
7.3.1	User Experience Questionnaire	32
7.3.2	Feedback from interviews.....	33
8	Evaluation of Epic 6.11: Intra- and interlingual subtitling.....	35
8.1	Motivation	35
8.2	User test setup	36
8.2.1	Material used.....	36

8.2.2	Participants	37
8.2.3	User data collection and tasks	38
8.3	Analysis of user data	40
8.3.1	Productivity data.....	40
8.3.2	User Experience Questionnaire	46
8.3.3	Feedback from interviews	49
9	Discussion and impact of the second prototype evaluation.....	53
9.1	Overall observations of the second evaluation round.....	53
9.1.1	Overall conclusions concerning the evaluation of “Editing assistance using multi-modal and multi-lingual metadata”	53
9.1.2	Overall conclusions concerning the evaluation of “Searching and browsing for ingested and archived content”	54
9.1.3	Overall conclusions concerning the evaluation of “Intra- and interlingual subtitling”	54
9.2	Impact on metadata usage and format specifications.....	55
9.3	Impact on the future evaluation of the prototype	55
9.4	Impact on the development of the final MeMAD integrated prototype.....	59
10	Future prototype development and evaluation plan and dissemination activities	63
10.1	Final project year evaluations	64
10.2	Dissemination activities	65
11	Bibliography	66
Appendix A	Epic 6.5 – Editing assistance using multi-modal and multi-lingual metadata: Evaluation tasks, participant editing script briefing, think-aloud instructions and post-evaluation interview.	67
Appendix B	Epic 6.2 – Searching and browsing for ingested and archived content: Evaluation tasks, participant briefing, post-evaluation interview and participant background information form.....	75
Appendix C	Epic 6.11 – Intra- and interlingual subtitling: Evaluation tasks, participant briefing, post-evaluation interview and participant background information form.	84
Appendix D	Participant introduction guide: Searching and browsing in the MeMAD prototype (Epic 6.2, User Stories 2.2.*)	92

1 Introduction

In this deliverable, the second of three iterations, we describe the results of the second evaluation round of the prototype MeMAD platform as described by D6.5 and performed as part of task T6.3. This deliverable reports on the user feedback and observations made by the consortium during the end user evaluations that took place at the end of the second year and the beginning of the final year of the project.

The second round of testing evaluated those components that were available in a mature form at the end of the project year, which 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;
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 MeMAD-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).

Section 4 explains exactly 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), thereby providing context to the overall project work plan and use cases that will eventually be implemented.

To provide the reader with additional background, we additionally summarize how this deliverable compares the previous evaluation iteration report (D6.1), in Section 2. Then, in Section 3, we also provide context regarding the place of the evaluations in the overall human-centred design and implementation process adopted for the development of the MeMAD prototype.

Section 5 describes the general evaluation methodologies that we adopted for performing the end user evaluations. Sections 6 through 8 each functional evaluation is then 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 amount each type of data collected (questionnaires, interviews, usage metrics, etc.).

In Section 9 we take a broader look at the evaluation results in order to determine their impact on the future work plan of the project, and this along three axis: the future requirements and exchange formats devised for the project, the implementation of the final MeMAD prototype, and any impact this evaluation has on the future evaluations to be executed in the final project year. We conclude this deliverable in Section 10, which

also defines the tentative evaluation calendar for 2020, the subject of the final iteration of this series of deliverables; D6.9.

2 Changes with regards to deliverable D6.3

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

1. The formal development methodology introduced in version 2.0 of D6.1 (and further refined in D6.4) has been adopted in Section 3 and formed the basis for continued work done in T6.3 from Month 13 to Month 24 in the project. As opposed to D6.3, this gave us a proper framework to determine the goals and context for each executed evaluation.
2. Following on the previous point, Section 4 was added in line with the corresponding section in D6.5 to pinpoint the scope of the evaluations and how they match the finished functionality of the prototype platform.
3. Section 5 introduces the process methodology used for this round of evaluations, which is an improved methodology compared to that used for the evaluations described in D6.3. We did incorporate the recommendations from D6.3 and D6.4 to properly define this round's methodology, and it is more fitting for end-user software evaluations as done in this round, as opposed to gathering mostly end user impressions (as was the case in Y1 of the project).
4. The "Interviews" Section from D6.3 has been entirely replaced by Sections 6 through 8, each of which is dedicated to a single functional evaluation.
5. The impact section 9 has been retained but extended; we now provide a general conclusion on each function evaluation, and then conclude the impact of this evaluation round on the same axis as before: the future requirements and exchange formats devised for the project, the implementation of the final MeMAD prototype, and any impact this evaluation has on the future evaluations to be executed in the final project year.
6. We have extended the final section with a tentative calendar of evaluations for the final project year.

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 as functional foundations for the prototype's development.**
2. **With regard to the Human-centred Design methodology¹ 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.
 - 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 of each iteration of the MeMAD integrated prototype. Deliverables D6.1 and D6.4 provide these guidelines.
 - The development of the prototype also falls under the 2nd phase of the UCD process, which has now completed its second cycle, based on the initial round of evaluations reported in D6.3 and the subsequent revised specifications from D6.4. The selection of D6.4 functionalities for the second iteration was done at two plenary consortium meetings, one at University of Helsinki in April 2019, and one at University of Surrey in September 2019 for a final selection and confirmation.
 - Finally, in Phase 3, the implemented second prototype platform iteration has been evaluated with end users, first of all 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. This work was performed as part of T6.3, and this deliverable is the report of this evaluation cycle.

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

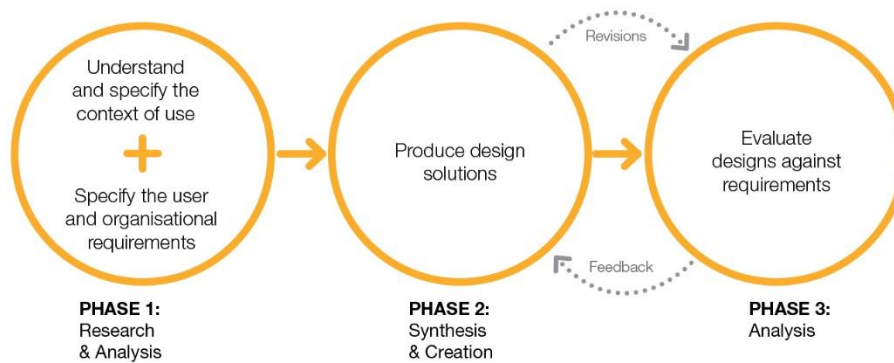


Figure 1: Synopsis of the User-Centered design process (from O’Grady, 2008²).

As explained above, because the UCD process encompasses the entire duration of the project, not all of its results are available in this deliverable yet. We summarize how each piece will subsequently be completed 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 (M36).	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 (M36).

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 that was evaluated in the second evaluation cycle of the project. This second iteration was guided strongly by the requirements defined in D6.4 and was focused much more on end-user functionalities than the first iteration was. While the implemented functionality of the platform is extensively described in D6.5, we summarize these functionalities here as a context for the describing the various evaluation tasks that were executed.

Considering various constraints based on the availability or maturity of services from WP2-WP5, the availability of end users to evaluate the prototype, the progress of research and insights on topics such as machine translation (MT), legacy metadata processing, narrative video captioning, and finally, based on feedback gathered from members of the project's External Collaborators Group, a selection of functionality was made to be implemented in the second prototype platform. Expressed according to the functional epics and user stories defined in D6.4, Epics 6.2, 6.3, 6.5, 6.7 and 6.11 were elected to be worked on for this version of the prototype platform (cf. highlighted boxes in Figure 2). Again, for a more elaborate derivation of this selection, we refer to D6.5.

With regard to the prototype evaluation, we summarize the functionality that was built as part of the prototype, and how it was evaluated in Table 2. An extensive discussion of each evaluation is then given in later sections of this deliverable.

We can summarize that the functionality of the second MeMAD platform integration has been evaluated in the following three separate groups of evaluations:

4. 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;
5. 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;
6. 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.

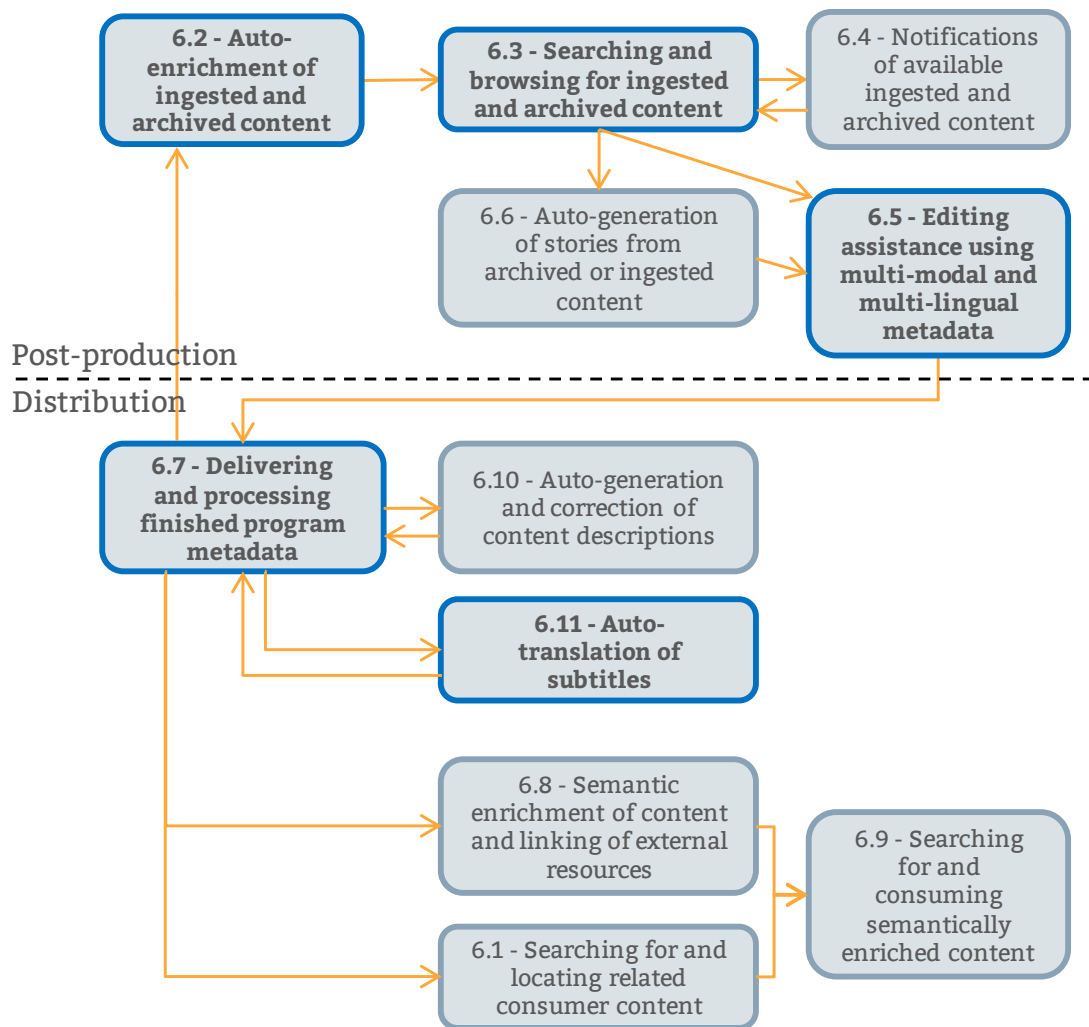


Figure 2: Functional epics selected to be implemented for the second MeMAD prototype platform.

Epic	Applicable User Stories	Implementation	Evaluation
Epic 6.2	2.1.1 - Real-time analysis and indexing of ingested content, 2.1.2 - Extensive analysis of ingested content.	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 iteration, the focus of integrations was on named entity recognition, automated speech recognition, face recognition and machine translation.	It is the outcome of the processing components that is used by end users for searches or as a starting point for authoring subtitles. As such, this functionality was evaluated as part of the evaluation of Epic 6.3, 6.5 and indirectly also for 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.	Idem.
Epic 6.3	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 by end users using the prototype's user interface, fueled by a combination of both legacy and newly auto-generated metadata. This evaluation track is discussed in Section 7.
	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). Selections and exports can also be made on a subclip level basis.	
	2.2.5 - When looking up archival content,	As part of the named entity disambiguation process, links to relevant knowledge bases	

	hyperlinked related media are also shown.	(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.	
Epic 6.5	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.	An evaluation track was executed to test the usability of providing this auto-generated metadata in dedicated video editing environments, commanded by professional video editors. This evaluation is discussed in Section 6.
	2.1.6 – Use of autotranslated 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.	
Epic 6.7	2.3.2 - Delivering relevant production metadata downstream.	A first set of proof-of-concept export mechanisms were implemented to allow the MeMAD platform to export production metadata to downstream production processes to external systems. These metadata include shot-cut-boundary information created by the platform's media processing services, and subtitles generated and manually authored within the platform.	This functionality was built to support MeMAD services which are under construction at the end of Y2 of the project, and are as such not yet being evaluated.
	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	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

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/>.

		stories implemented by the MeMAD platform.	indirectly evaluated in Section 7.
Epic 6.11	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 evaluated with professional subtitlers, as elaborated in Section 8.
	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.	Interlingual subtitling, encompassing both the generation and post-editing of the subtitle and subsequent qualitative evaluations are discussed in Section 8.
	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.	

Table 2: Functional epics and user stories implemented in the second MeMAD platform prototype.

5 General evaluation setup and methodology

This section describes the common approaches and methods utilized for the evaluation of three MeMAD epics described above in the previous section. The aim of the evaluation is to understand the usability of the MeMAD technologies in metadata creation from the user perspective; thus, evaluation is not based on automatic metrics but on the analysis of user experiences. In this second 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, the second evaluation round took a “bottom-up” approach to the study of usability and set up a study which yields insight into the perceptions, attitudes and opinions of users towards the new technology. Based on the knowledge gained from this second iteration, we can plan and execute the final, third iteration with precise scenarios for testing the technologies and their impact to working conditions and productivity.

The first round of evaluations was based purely on interviews of media professionals (cf. D6.3). In this second round of evaluations we gave the participants a more hands-on experience, building the test situations 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 three 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 apply 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.

In this second iteration of the evaluation, we combine qualitative and more quantitative approaches and gather data from users performing controlled tasks with the following methods: Think-Aloud Protocols or process data (keylogging) during the tasks, and User Experience Questionnaire and Semi-Structured Interviews after the tasks were completed. These methods produce data on subjective evaluations by the users; their impressions, feelings, opinions, and attitudes, as explained in the following subsections 5.1 through 5.3. To the extent possible, we supplemented these subjective assessments with objective evaluations, for example by quantifying user assessments with scores obtained from the User Experience Questionnaire and actual timing and keystroke measurements when testing the subtitling processes, as explained in subsection 5.4.

5.1 User Experience Questionnaire (in all cases)

For all three use cases in this second 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 the purposes of this evaluation, a modified version of the UEQ was used. For all use cases, 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 a 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 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.

5.2 Semi-structured interviews (in all cases)

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.

5.3 Think-aloud Protocols (for Epics 6.3 and 6.5)

In the evaluation of user stories 2.1.* (on video editing assistance) and 2.2.* (on searching and browsing content), we applied think-aloud protocols [4] during the tasks in order to obtain more fine-grained information about how the participants approached the tasks and what potential problems they encountered and what kind of solutions and decisions they made. Thinking aloud as a research method originates in psychology but is

nowadays widely used in various disciplines. It is widely applied in usability studies [5]. The purpose of a think-aloud protocol is to elicit data from the participant regarding their processing of a task: what they were thinking, what potential problems they encountered, how they solved the problems. Following the guidelines given by [6], the participants were instructed to verbalize their thoughts out loud as they carried out the editing or search tasks. The verbalizations were recorded and transcribed for further analysis. The analysis of the think-aloud data is still on-going, and results will be reported at a later stage of the project.

5.4 Process data collection (for Epic 6.11)

For the evaluation of user stories involving intra- and interlingual subtitling, 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. Keylogging software (Inputlog, see [7]) was used to collect process data during the subtitling tasks carried out by the participants. Process data (task time and number of keystrokes) were then analyzed comparing tasks where the participants post-edited ASR or MT output to tasks where they created intra- or interlingual subtitles from scratch. Collection and analysis of process data is described in more detail in Section 8.2. For further discussion of MT post-editing in user story 4.3.4, see also D4.2.

5.5 Materials used in the evaluation

For evaluation purposes a subset was selected with the theme of European Parliament Elections from the MeMAD dataset provided by project partners Yle and INA. 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.

Main program genres used were journalistic factual and news programs because these are the main genres of all the MeMAD media datasets. Through this genre selection the evaluation tasks focused on task types typical for news, factual and documentary productions, they did not cover e.g., task types more common in fiction or drama productions.

For the searching use case 2.2 the whole subset of 408 tv and radio programs (209.6 hours total) was used, including ancillary materials such as pre-existing subtitles and legacy metadata provided in the project datasets. For the video and subtitles production use cases a small subset was selected as the time spent on these tasks depends closely on the amount of content handled. 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 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). The purpose of collecting and analyzing process 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.

A process study pilot was carried out in December 2019 at the YLE premises. In this study, professional video editors edited a short (c.a. 2 minutes) video mash-up from a longer set of raw footage using automated metadata, ASR output and MT output to support the searching, browsing and editing of the video content.

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 5.1 and 5.2). During the editing tasks, think-aloud verbalizations were also collected from the participants (as introduced in subsection 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 availability of auto-generated metadata. The task evaluated is the video editing process in which users assemble a new program (item) from rough parts. Additionally, we want to assess the extent that editors can be helped by machine-translated speech transcripts for understanding other-language content during the editing process. The test panel worked within their expert editing software environment to ensure a representable implementation of the task.

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.

6.2.1 Material used and testing environment

From the MeMAD catalogue (cf. D6.5 and D1.2) a single video was chosen to be used as raw footage for the editing case, 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 elections⁵. 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.

For this 2-hour clip, we processed the clip using the various MeMAD components to make the following set of metadata available to the video editors (cf. also D6.5 on more details on the edit suite integration work):

- The source audio material was auto-transcribed from English, Finnish and French speech. This was done using both ASR components delivered within the consortium, as using commercially available solutions. The Finnish speech was processed using Lingsoft's ASR, as described in D2.2. The English and French speech fragments were analyzed by the Speechmatics ASR service in the version publicly available in November and December 2019⁶. We ensured that the transcript parts were manually cut up into manageable sections of at most a couple of sentences, where applicable. This manual intervention was done to make sure the data could be exchanged with the edit station, and to avoid overly long transcript lines in the Avid editing application (as there is no way to make their visualization multi-line). In this process, the content of the ASR transcripts was not modified.
- Non-English transcript parts were machine translated into English. As with the ASR execution, this process was split amount commercially available service for highly-resourced languages incl. English and French and MeMAD-specific developments provided from the consortium for the Finnish language parts. For English and French MT, the DeepL translations service was used, again in the version available in November and December of 2019⁷. The Finnish MT software was developed at University of Helsinki and is described in more detail in D4.2.
- Face recognition was performed on the participants of the video content, as provided by EURECOM and detailed in D2.2. The recognition service was specially trained on the set of faces present in the video material. The results were made available as timed textual metadata.
- Named Entity Recognition (NER) was executed on the same-language transcripts delivered by the ASR components. NER results were grouped per transcript paragraph in which they were detected, and then also made available as timed textual metadata. NER was executed using a variety of NER extraction services from the consortium, incl. from EURECOM for analyzing French transcripts and from Lingsoft for handling the Finnish transcript parts. The implementation of both services has been described in D3.2. In addition, a commercial service, in the version available in November and December 2019, from TextRazor was used for deducing named entities in the English transcript sections⁸.

⁵ Also available online at: : <https://areena.yle.fi/1-50141056>

⁶ Cf. <https://www.speechmatics.com/product/features/>; we used the Software-as-a-Service (SaaS) Batch ASR processing service.

⁷ Cf. <https://www.deepl.com/en/docs-api/>; we used the v1 of the DeepL translation API.

⁸ Cf. <https://www.textrazor.com/docs/rest>; we used the SaaS REST version of the text analysis API, and specifically, the results returned from the "entities" extractor.

- Finally, legacy metadata from the YLE archive system was also included in the metadata for the editing task (cf. D6.5, D3.1 and D3.2). Both temporal and non-temporal data was included, e.g., the genre, topics and segmentation done by archivists on the item.

The following images depict examples of finalized data that was exported to the Avid editing environment. The selected audiovisual clip is shown in the background, along with indications for all temporal metadata markers (shown as red dots) along the clip's timeline shown in Figure 3. An actual breakdown of metadata stored under the markers list view is depicted in Figure 4.

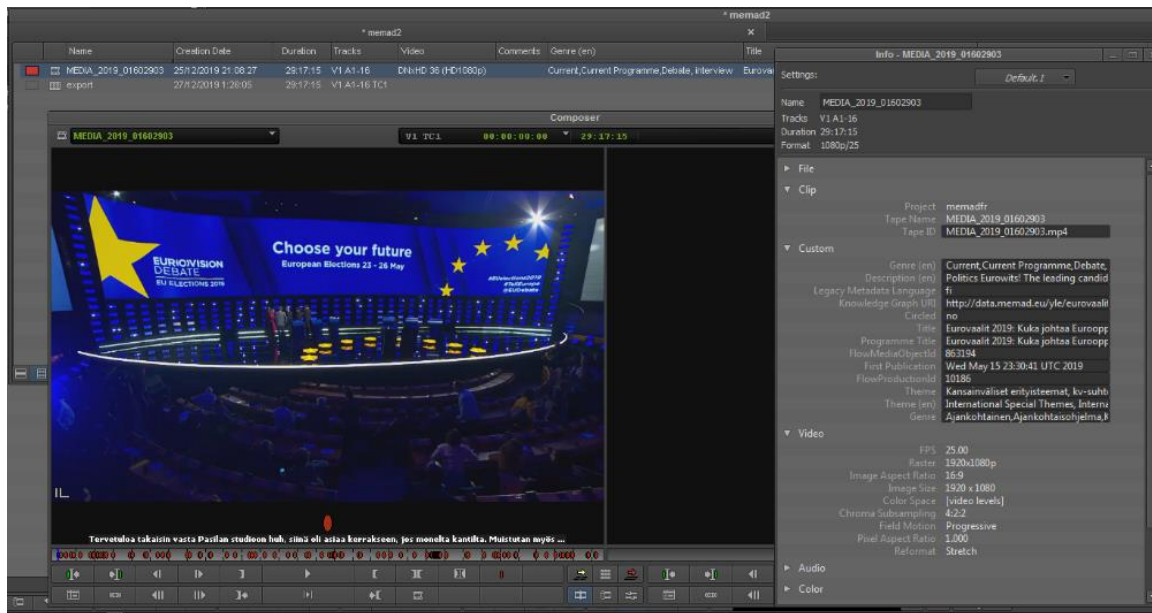


Figure 3: View of the final set of program metadata imported into Avid Media Composer. The MeMAD metadata is shown as custom fields in the clip Info window.

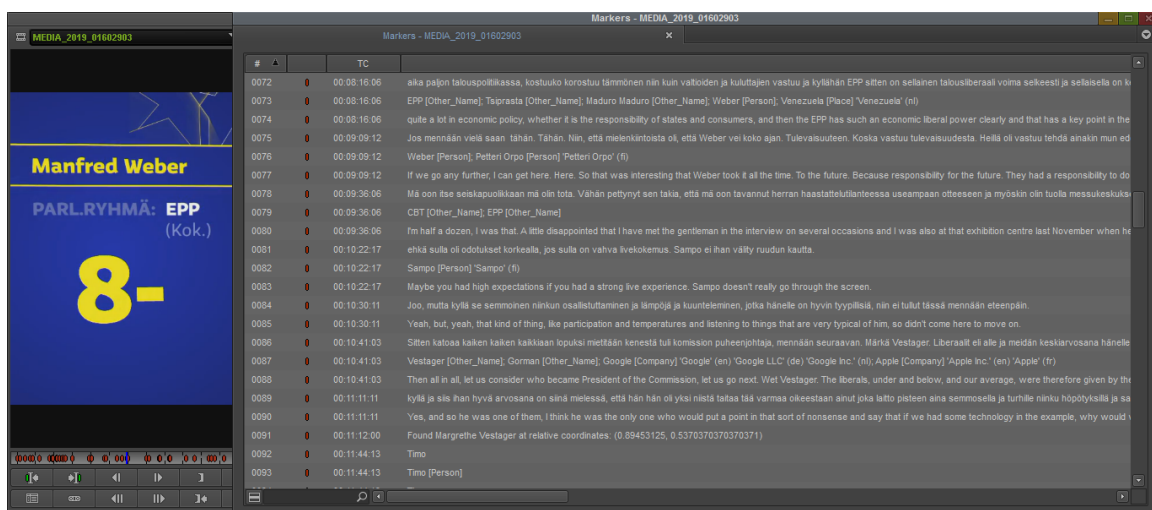


Figure 4: View of the final set of segment metadata imported into Avid Media Composer. The metadata is displayed as markers on the clip timeline and in the Markers window.

6.2.2 Participants

Three professional video editors were recruited as participants. Participants in this round of testing were in-house video editors working for the project partner YLE. The Avid Media Composer professional craft video editing software was chosen as the evaluation tool as it is the software used in-house for video editing at YLE, and as such, 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.

6.2.3 User data collection and tasks

The experiments for editing assistance data collection were arranged at YLE premises in December 2019. As mentioned above, the editing assistance tasks were carried out using Avid Media Composer. The editors had access to the internet and most other resources normally used in their work. One out of the three editors set up their own shortcuts and key-bindings before starting the test.

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 editing experience and the quality of the available metadata. Additionally, a screen recording of the editing process was made with screen recording software provided as part of Windows 10. 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 participants were instructed to produce a video containing all the specified segments, and to use any resources they normally would for their work (e.g. the internet), but not to spend excessive time on “polishing” the video or making it exactly match the given length of two minutes. No explicit time limit was given for the tasks, rather, the participants were instructed to work at their own pace. The list of tasks given to participants can be found in Appendix A.

6.3 Analysis of user data

6.3.1 User Experience Questionnaire

In UC2.1, the UEQ questionnaire was used after the editing task to collect the participants' subjective evaluations of the experience of using metadata for editing. Figure 5 shows the averages of the three participants' responses to each of the questions on a scale from -3 to +3. On average, the responses tend to be mildly positive or neutral.

The most positive responses characterize editing with metadata as relatively exciting, fun, motivating, pleasant, enjoyable and efficient. When asked specifically about the quality of the automatic speech recognition, machine translation and face recognition, the participants' reactions are slightly negative. For named entity recognition, the response is neutral, suggesting that this type of metadata appears to have worked better than the other types although on average it is not characterized positively.

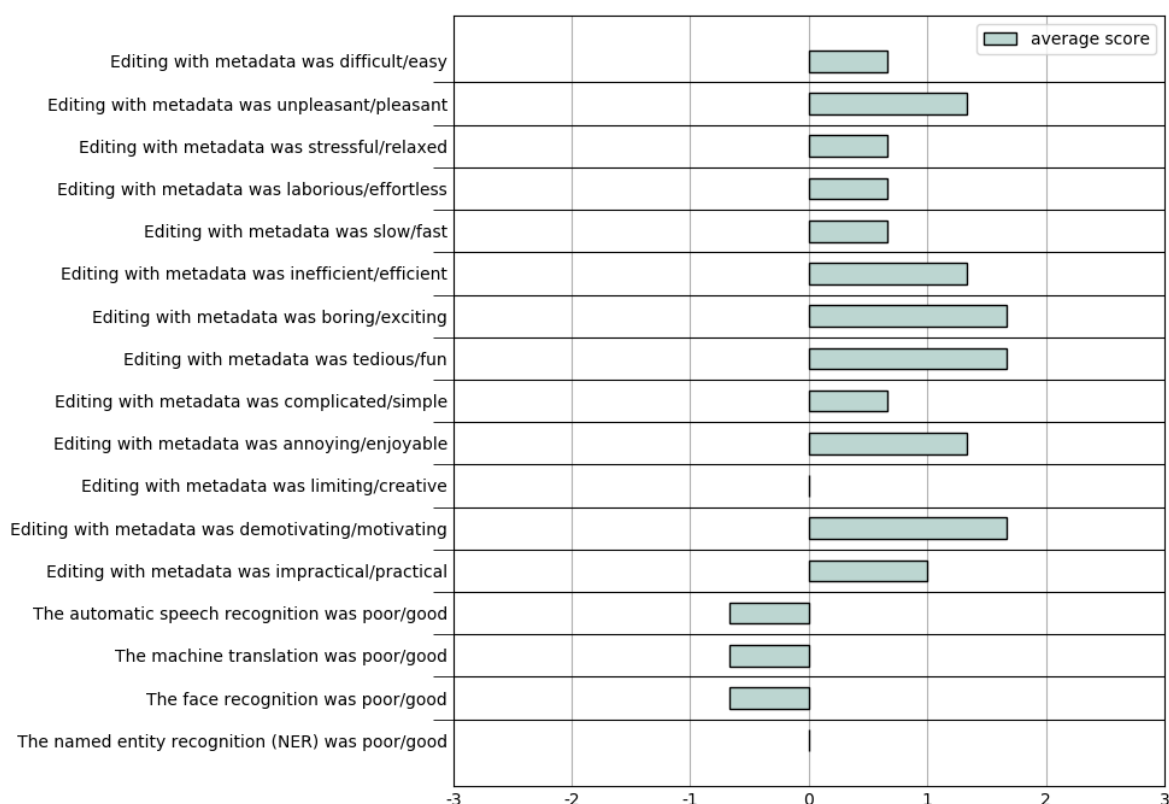


Figure 5: Average UEQ assessment scores for the participants in the video editing assistance evaluation.

6.3.2 Feedback from interviews

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

1. How did the editing task feel overall? (Optionally, if the participant's overall answer was negative, ask about positives and vice versa.: Was there anything positive/negative about the task?)
2. What features of the metadata impacted the editing most?
3. Did you notice any differences in the quality of the metadata or transcriptions?
4. What is your editing process usually like with this kind of material?
5. How did the use of ASR/MT impact your own work process?
6. What about the use of NER/face recognition, did you use them/how did it impact your work?

7. Could you imagine using ASR/MT for editing?
8. How should this [metadata, not the interface] be improved as a tool?

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. The number of interviewees being limited ($n = 3$), we did not conduct quantitative analysis of this interview, as we did with the subtitle post-editing task (cf. Section 8).

Overall, both positive and negative issues were raised. The possibility of searching and finding material via metadata was perceived useful in general but it was considered not to be an essential part of editing. Two out of 3 of the participants regarded that the metadata functionalities tested (ASR, MT, NER and face recognition) are more useful for journalistic work processes than for pure video editing. One was quite critical about the task and claimed it unrealistic. This was explained by the editors recruited for this task; as they are typically tasked with editing tasks that are accompanied by precise editing instructions from journalists, directors or producers. As such, decisions on which descriptive metadata would have a large impact would have already been taken by the time the editing task arrives at the pure video editors.

On the other hand, Two out of 3 had a generally positive feeling about using metadata, yet one of them commented about feeling stress in the situation.

On the negative side, the participants referred to particular issues in using the technology and in benefitting from the output of that technology. 2 participants referred to integration and performance problems and in the combination of two software used (Avid for editing and Flow as material database). One participant also regretted that s/he uses different logic than the interface but added that the two (user and interface logics) should work seamlessly together. Indeed, the coupling of user's logic with the metadata representation logic was mentioned by all participants as a problematic issue. The way in which users verbalize their thoughts may be different from the metadata format: e.g. search terms used do not always match with metadata wording. One participant regretted that only words, not phrases, can be used for searching in the database of the craft editing software. In addition, the participants suggested some improvements for the form of representing metadata. Instead of increasing quantity, more precise metadata are needed: e.g. matching text with corresponding shot in the media, and highlighting search term in the text. Also, the format in which two different Avid markers were given for face/person and speech/text is not helpful. Similarly, due to the way Avid displays temporal metadata as markers, sections of automated transcripts are always shown a single line, which becomes cumbersome for larger paragraphs, or even longer sentences, of transcribed dialogue.

On the positive side were aspects about the general work organization, the use of translations, and the general positive, enthusiastic attitude about these new technologies. The participants acknowledged that the metadata helps searching and processing audiovisual material and the quality of the output is good enough to deal with fast work processes. Among the perceived advantages was the ability of computational methods to ease the manual work and thus reduce workload and process material more

rapidly. Two of the 3 of the participants could imagine using ASR and MT for their work, and the third one considered ASR to be useful for “lazy journalists”. Furthermore, another 2 out of 3 participants considered that searching content via translated speech is useful, although one participant reminded of the risk of errors in machine translations. This relates to another type of risk that was mentioned: in using automatic technology, the perceived ease of working (e.g. processing material) may produce false self-confidence which then leads to errors.

One functionality, face recognition, evoked mixed opinions. One participant had used it and considered it helpful (“to match name with frame”), while another one did not and used Google search, “the traditional method”, instead. This one, however, also regarded face recognition as a useful tool in general. Since we did not explicitly ask about the usefulness of face recognition or NER for editing, we did not get concrete insight from the participants about including those in the future development of the metadata functionalities.

7 Evaluation of Epic 6.2:

Searching and browsing for ingested and archived content

In this year's evaluation round, 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).

A process study pilot was carried out in January 2020 at the premises of YLE, and at the premises of the Finnish National Audiovisual Institute KAVI. In this study, professional media archive users performed a series of media and information retrieval tasks using different combinations of automated metadata, ASR outputs and MT outputs.

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.

Archive users' subjective evaluations of the usability of auto-generated metadata, ASR and MT for these purposes were collected using the UEQ survey and semi-structured interviews. During the searching tasks, think-aloud verbalizations were also collected from the participants.

7.1 Motivation

The aim of this evaluation is 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.

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.

7.2.1 Material used

From the entire collection of audiovisual media contributed by YLE and INA, a subset of 408 relevant items (amounting to 210 hours of content) were selected by representatives of YLE and INA, a selection of various materials, incl. current affairs, lifestyle, politics and new broadcasts. This subset of the entire MeMAD catalogue of content was made available through a separate sub-collection in the Limecraft Flow prototype user interface.

With respect to metadata enrichments, the following data was made available for the end-user testing:

- Items had their ‘legacy metadata’ (cf. D6.5, D3.1 and D3.2) available from the original archive system at YLE or INA (through the MeMAD Knowledge graph, cf. D3.2). This metadata was stored as subclips for the original segments, with an additional subclip to represent the overall description of the clip. Other fields of legacy metadata were available as clip metadata in the info panel. These include:
 - Program and episode title;
 - Genre;
 - Themes;
 - Working title (if applicable);
 - Date of first publication/broadcast.

Each piece of legacy metadata was machine translated into English.

Figure 7 illustrates the user’s view of the legacy metadata in the MeMAD platform.

- All but a few items were audio-transcribed by MeMAD services. This includes items that have the majority of speech in Finnish, Swedish, English and French. Clips left out of the ASR chain were those in Northern Sami (for which no ASR was available), and highly-mixed language items for which the choice of a single language ASR engine would produce unusable results for all other language segments. All transcripts were directly sourced from the available ASR services and no post-editing was performed on these transcripts. The ASR processes were executed in the same way as described in Section 6.2.1.
- All items with speech transcripts in French or Finnish have been machine-translated into English as the common language. Lower-resource-language

transcripts such as Finnish were translated to English using MT tools made available within the consortium, while other items, in particular those in French, were translated by tools from commercial providers (incl. DeepL), again, as also described in Section 6.2.1. Note that the legacy metadata was translated from its source language in the same fashion.

- All items with an audio transcript (in any source language) were processed by a named entity extraction service, for which Section 6.2.1 also provides details.

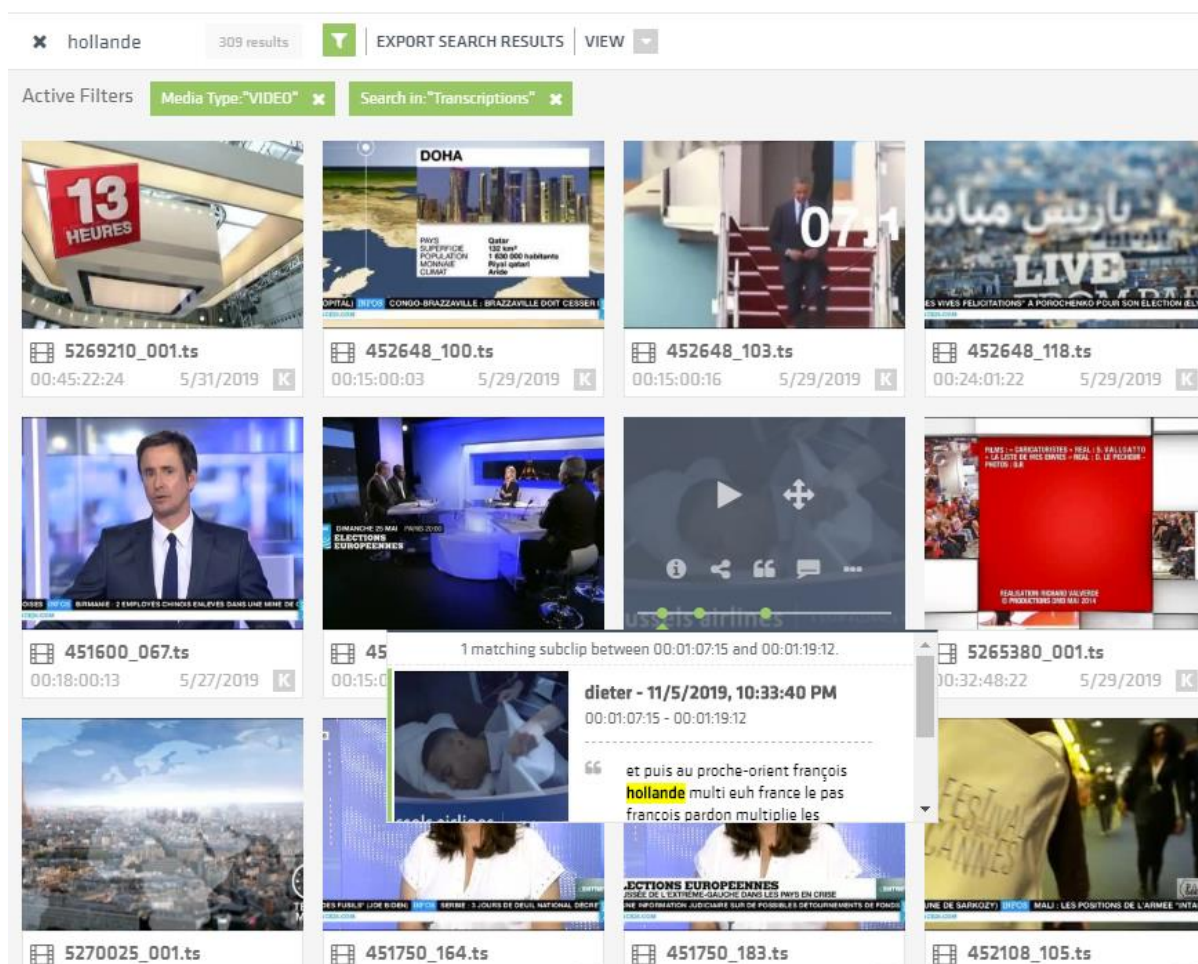


Figure 6: The platform library shows search results for the term "hollande", with clip and audio transcript part matches in the search results.

To illustrate the way ASR and NER data was presented to the test panel, we have included Figure 8. In addition to the legacy metadata shown above, custom-made views for speech transcripts and named entities were available to the end users of the platform.

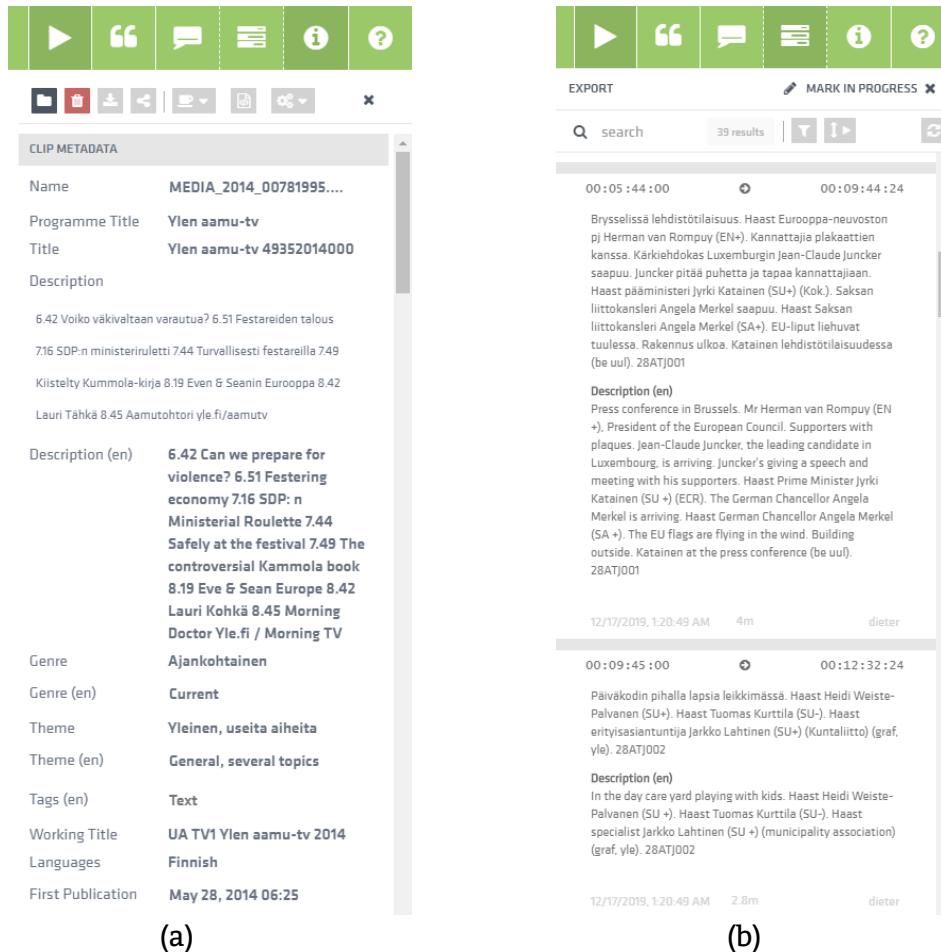


Figure 7: Final MeMAD platform views of imported legacy metadata.

(a) Program-level metadata and (b) segmented metadata, each with translations for all metadata fields.

We refer to Appendix D, which provides 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

The screenshot displays the MeMAD platform interface. On the left is a video player showing a news broadcast. In the center is a speech transcript in Finnish. On the right is a table of Named Entity Recognition (NER) results.

Speech Transcript (Finnish):

maksetaan niin paljon parempaa palkkaa sinä väneessä kuin rannelausuppoa yläpäässä ja ratkaisuja. Hienoa. Kukaan muukaan ei välttämättä joutunut poikkeavien mielialojen, mutta sinä olet populaarisesti käynyt, johtajien pakojen kiinnittämisen, miten sinä vastaat Apuella, en ole tehnyt sitä populaarisuutta Matti Apuella on yleensä argumentit koulusta, mutta täällä kiertää. Ne on kyllä kevyttä kamua ensimmäinen omistaja päättää, jee, tosiaan. Suomessa yritysjohdajien palkkoista päättää muut yritysjohdajat, jotka istuu hallituksissa istuu hallituksissa kuka siellä istuu porssiä ja hallituksen entisiä nykyisiä johtajia tai yhtäkkiä kokouksessa ei ne sadat pöytäkirjoista siellä mitään päätetään on eläkeehtojen ja muiden edustajat, jotka näitten yritysjohdajia, jolla on intressi siihen, että mahdollisimman kannustavat sytyneet olo toisiksi hän puhuu maailmanlaajuisia suomalaisia IHL, kielikolijat on maailman huppu, onko suomalaiset yritysjohdajat maailmanhuppu, jos on, miksi näitä rekrytoitua tuonne Wall Streetille. Missä maksetaan moninkertaisesti vielä se, mitä suomalaiselle yritysjohdajalle Suomessa pöytäkirja on se, et yhäkkytykseltä alkaa, jolla ei ollut tätä digitalisaatiota mistä hän nyt puhe yhdeksän levin aikaa alkaen suomalaiset yritysjohdajat on tehty merkittäviä liittoja koko muussa suomalaisessa yhteiskunnassa korkeista virkamiehistä Matti hetemä ja muut tekee huikeita duunia maailmanlaajuisen akateemikoista tutkijasta pääministeristä keskeisverroksista puhumattakaan mitään tämä liitto perustuu, ei mikään, ei muu mun mielestä Ben ei oo missään määrin tässä populaarisesti toiminut, että mun mielestä enemminkin erittäin johdonmukaisesti nostan että kyllä haastaa tää nimu mun mielestä seuraliitit tää tämän tarkoituksen yksi, 2 013 Missä tota toimittaja Jari Korvola kävi läpi näitä suomalaisia haapajohdajia, kunlka paljon näitä löytyy maailmalta. Kun IHL:stä näin sanotusta, jos tätä Patrik Laine vertausta käyttää, niin siellä kyllä montaa oo et ei ne vaan ei näytä tehneet vaan rita sinne, että kun mitään ymmärtää, et ne on SH-läppäseläjä näin, että ensimmäinen minusta halpaa se, että joku näistä palkkoista koko ajan puhutaan sitten kun media nostaa näiden ikään kuin aiheus saa kaavot ja nostin sinne joku yritysjohdajia ja katotaan, että hän edustaa jotain moraalista tapaa, kun hallitsee on pätkittänyt näin palkkansa ja hän ottaa sen mikä minäkin otaisin ja otaisin vielä enemmän, jos saisi koi Teos on mitään populaarisuutta, että aiheudelle annetaan mukaan kannot, yritysjohdajien palkkakeskustelun kautta. Mutta millä tavalla se sitten aiheutuu, että ennen tätä pain, tai olo ma on tää jengä, joka on sitä mieltä, että että valtiotyövoimien palkat tai valtiotyövoimien palkat on tää nimu sinne, että todella illan samat kuin kuin mitään mitä yksityisillä sektoreilla, ettei nyt voi olla kovin vaikeata myydä monopolisematta viinaa suomalaisille. Ainakaan suomalaisille. Kyllä voit tietää, että tota ja jaa sillo, koska ne on niin korkeat sen takia, että yksityisen sektorin ne on niin korkeat ja yksityisillä sektoreilla niin korkeita, koska, koska näin on jossain väheessä on tapahtunut ilmeisesti osinkojen osuuden osuuden optoiden tota jälkeen, mutta tämä tekoo sillo mahdollistamaksen muihin miehiä sitten kritisoita se näin kun Ruususen mielestä varmaan ketään ei voi kritisoida, mutta kun pensalta silloin aikaisemmin se, että siellä on noin Hesarin selvitäksen mukaan muutama kymmenen henkilöä, jotka Suomessa istuu suurimmissa osuista porssiähtiön hallituksista ja toimintajohdajana istuu ritin ritin paikkaa tuosta lähtien. Silloin sinne ei voi puhua kyllä suoraan siitä, et jos omistaja. Yksikään ole halpaa sillo antaa jokenkin aina puhutaan tästä niin sanotusta aiheudesta, psykologioissaan markkinatutkijasta, että väitetään, että näillä yritysjohdajilla olo nyt jotakin niinku luono moraal, kun he ottaa sen, mikä on saatavissa. He kaikki muuttii mä en ole milloinkaan lähtenyt siitä, että yritysjohdajat on se aine ja nollin ajatellen lähtökohta, kuin muutakaan totta kai voi ajatella et joku, joka menee valikka diakoniasaksi, niin ehkä raha ei ole kellekään yhtä tärkeää kuin semmonen, joka on käynyt kauppa- ja teollisuuden ministerinä. Totta kai Patrik on edustanut jossain määrin sen oloa. Yksikään

NER Results Table:

LABEL	TYPE	TIME RANGE	MORE INFO
Suomi	Place	[00:28:00:02,00:28:31:00]	view data
Vantaa	Place	[00:28:16:24,00:28:17:00]	view data
Yhdysvallat	Place	[00:36:16:01,00:36:18:00]	view data
Neuvostoliitto	Place	[00:32:19:14,00:32:20:05]	view data
Ukraina	Place	[00:38:07:05,00:38:07:22]	view data
Irak	Place	[00:38:06:10,00:38:07:00]	view data
Kiina	Place	[00:38:08:05,00:38:09:00]	view data
Kansallistieteen	Place	[00:31:19:05,00:31:20:07]	view data
Moskova	Place	[00:31:34:34,00:31:34:34]	view data
Etelä-Viro	Place	[00:35:00:11,00:35:11:01]	view data
Itä-Eurooppa	Place	[00:35:06:13,00:35:07:01]	view data
Leningrad	Place	[00:31:13:13,00:31:13:13]	view data
Oulu	Place	[00:41:00:03,00:41:00:10]	view data

Figure 8: Detailed content item metadata with the video player on the left, the speech transcript (available in multiple languages), in the middle and the NER results in the right panel.

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.

To ease the use of this selection, we preconfigured four “quick views” through which users obtain a correctly configured search environment (which search results limited to the content selected for this user story evaluation, and limited to the intended metadata):

- “2.2 – Only Legacy Metadata”: filters down search space to only the legacy metadata of the clip⁹.
- “2.2 – ASR and NER”: filters down search space to the audio transcripts and NER results, i.e., the metadata that could be obtained from the audio signal in the audiovisual material.
- “2.2 – ASR and NER and MT”: filters down search space to the audio transcripts, NER results and the machine translation of the original language transcript in English. I.e., this is all metadata that was generated automatically, either directly from the source material (first order metadata), or in second order derived from first-order metadata.
- “2.2 – All Metadata”: does not filter the search space and allows users to look through all available metadata, as a combination of legacy and auto-generated metadata.

⁹ Even though the metadata fields *program title*, *episode title*, *genre* and *theme* are legacy metadata, they are available in each of the scenarios.

7.2.2 Participants

Professional journalists and archivists were recruited as participants. Participants in this round of testing were in-house archivists and journalists for the project partner YLE and media archivists from the Finnish National Audiovisual Institute KAVI.

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. Likewise, three participants were fluent in Swedish; the other two had to rely on machine translations when going through materials in Swedish.

7.2.3 User data collection and tasks

The evaluations for this functional epic were carried out in January 2020 at the premises of YLE, and at the premises of the Finnish National Audiovisual Institute KAVI. The test panel of professional media archive users performed a series of media and information retrieval tasks using different combinations of automated metadata, ASR outputs and MT outputs, all from the MeMAD integrated platforms's search user interface, based on the Limecraft Flow platform. 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 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 encountered during the search process and the participants' views on potential improvements.

Task summary

A set of six tasks was presented to the testing panel. Each task involves looking up items in the MeMAD content catalogue through the Limecraft Flow search interface. The tasks were devised in such a way to reflect actual media archive use and to avoid biasing tasks towards a particular project partner's routine, they were drafted together by INA and YLE, with feedback from other consortium members. We summarize list each of the tasks in the table below.

Task	Short title	Task Instruction	Evaluation
1	Specific program	"Find a program where François Hollande gives a speech (official speech at his office room)."	Basic search for a specific program, in its entirety.
2	Program type	"Find programs that are local / national election debates from the 2014 European Parliament elections; 2 programs from Finland and 2 from France."	Basic search on program type, resulting in the selection of a set of entire programs.
3	Topic	"Find 3 programs / clips where wind power is discussed".	Searches of programs or clips (segments of programs) that discuss a given topic.

4	Person + Topic	"Find where British politician Daniel Hannan talks about immigration."	Searches clips in which a given person is talking about a topic.
5	Location	"Find a debate/discussion shot in an outside setting in Moscow."	Searches clips from a given event, time period or place.
6	Object	"Find 2 programs/clips with horses appearing."	Searches clips in which objects and places appear.

Table 3: Content retrieval task summary.

As a baseline test, each task was trialed by YLE and INA experts actively working within the MeMAD project, to ensure that achievable tasks were proposed to the evaluation panel. Tasks 1-4 tasks were deemed achievable, while tasks 5 and 6 are much more difficult given the limited set of visually-oriented metadata available at this stage of the prototype implementation. They were included nonetheless to get obtain a benchmark value that can be improved upon later, but such that the overall set of tasks remains stable.

7.3 Analysis of user data

7.3.1 User Experience Questionnaire

In UC2.2 the UEQ questionnaire was used after each search task (tasks 1-6, see Appendix B) to collect the participants' subjective evaluation of the usability of the metadata for video searching. Figure 9 shows the average of the 5 participants responses after completing each type of search task on a scale from -3 to +3. Averages for each task are shown as separate bars from 1 (top) to 6 (bottom). Overall characterizations of searching with metadata appear to be positive particularly in the case of task 3 (light green bar), and to a lesser degree in the case of task 6 (purple). In contrast, characterizations of task 2 (orange) and task 4 (dark red), except for the adjective pair limiting/creative, where the responses for all tasks are on average positive, and task 2 is also characterized as exciting. Tasks 1 (light blue) and 6 (dark green) receive more neutral and mixed assessments, being characterized as slightly difficult, complicated, laborious and inefficient, but also relaxed, fun and motivating. When asked about the quality of the automatic speech recognition, machine translation and named entity recognitions, responses are mostly neutral. For task 3 and task 6, automatic speech recognition receives slightly negative assessments. Named entity recognition, on the other hand, receives slightly positive assessments in task 4 and task 1.

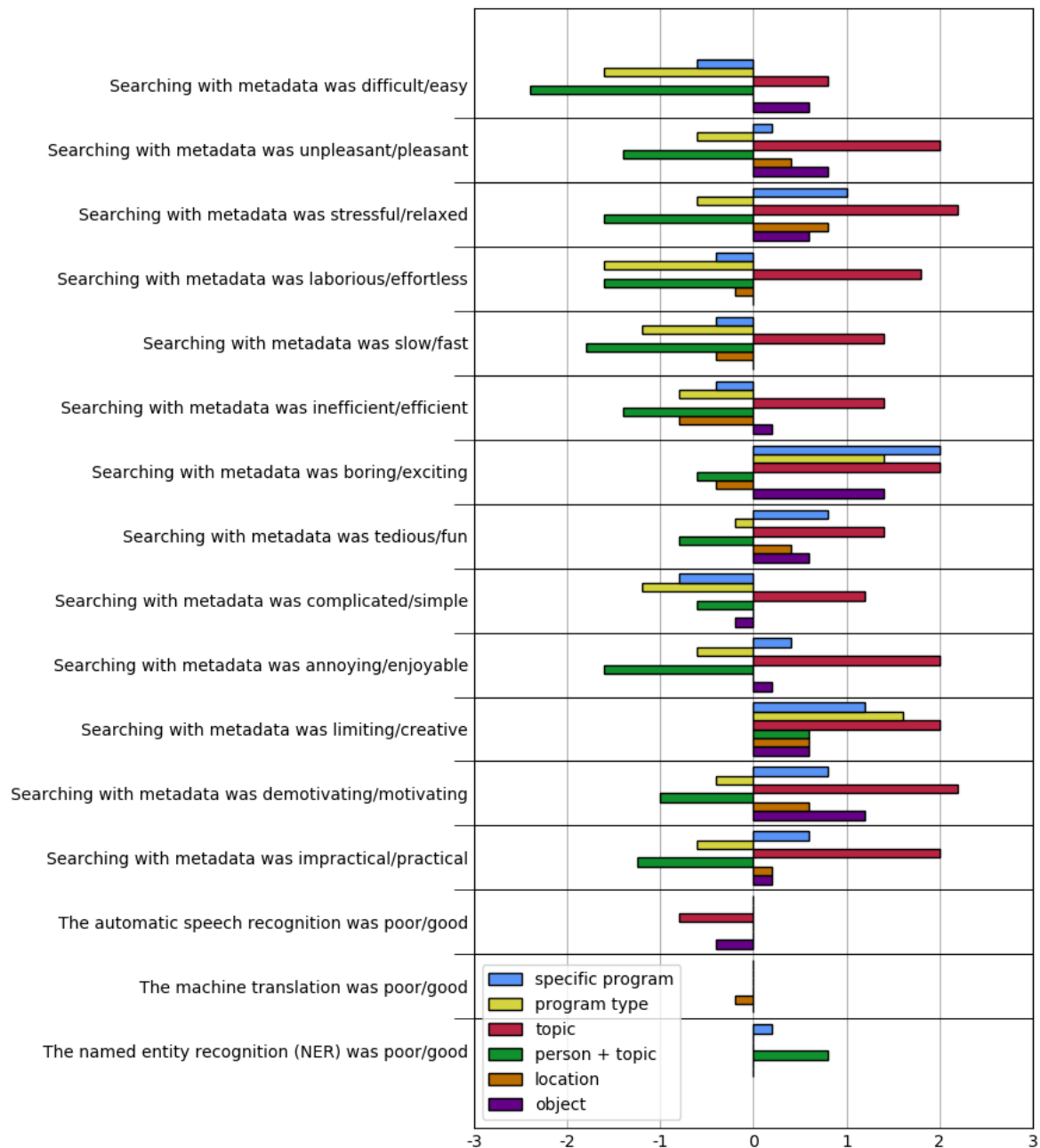


Figure 9: Average UEQ assessment scores for the participants when searching for content in the MemAD prototype platform..

7.3.2 Feedback from interviews

For the evaluation interview for user story 2.2.*, we adapted the questions (all but one) and structure of the interview in the following way:

1. How did the search tasks feel overall? (Optionally, if the participant's overall answer was negative, ask about positives and vice versa.: Was there anything positive/negative about the task?)
2. What features of the metadata impacted the searching most?

3. Did you notice any differences in the quality of the metadata or transcriptions?
4. How did these automatic metadata impact your work?
5. Could you imagine using this kind of metadata for your work?
1. How should this [metadata, not the interface] be improved as a tool?

Like in the evaluation described in Section 6 and in order to understand the user experience (using MeMAD metadata functionalities in video search tasks), the interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements. The number of interviewees was 5 (2 participants from YLE and 3 participants from Finnish National Audiovisual Institute).

Overall, both positive and negative issues were raised. The tasks were deemed rather difficult, and several participants complained about not performing well because they did not know how the system works (e.g. what kind of search terms should be used). In contrast, however, all participants expressed interest in using these functionalities in their work and they were also generally positive and interested in future possibilities.

Among the negative experiences the participants highlighted their inability to search correctly in the metadata: either they didn't know how to search (e.g. what terms or words to use) or they found "wrong" information (instead of visual content, which they are used to dealing with, they might find an expression of 'horse' in a metaphor in speech). Basically, however, they were content with the rapidity and ease of finding material as long as the search (term) was successful. Two important missing features were mentioned: the possibility of refining search (e.g. to certain categories, like 'shooting locations') and of filtering results (again, some categorization of results so that their relevance could be judged, e.g. when the 'horse' appears in speech or image). The participants also claimed that they didn't have time to judge the quality of the metadata due to the difficulty of the tasks. One participant wondered whether some searches were unsuccessful because of poor data quality.

On the positive side, the participants mentioned the quantity of data as one advantage, although some also warned of "getting lost" in the data. The quantity makes the searching more versatile and gives more results and therefore possibilities. Yet one participant also mentioned the importance of filtering the results if one wants to save time with this system.

One functionality evoked mixed opinions. Two out of 5 participants were perplexed about the text format of representing the data but, on the other hand, three of the participants considered the text format handy because they can merely scan through the data without having to watch the material.

As for improvements, the participants had several ideas. Two highlighted the importance of including image recognition and visual analysis to the metadata, perhaps even intelligent content analysis (e.g. indicating where 'horse' is relevant). The visualization of the metadata, similar functionalities as in other metadata databases and the use of standard vocabularies and even multilingual terminology in the metadata were mentioned as ideas for improvement. Finally, two from the 5 of the participants didn't think they'd know enough about the system to suggest improvements.

8 Evaluation of Epic 6.11: Intra- and interlingual subtitling

In the final part of this evaluation round, process data were collected from subtitlers working with intralingual subtitles created by ASR (user story 4.1.4), and interlingual subtitles created by MT from pre-existing intralingual subtitles (user stories 4.3.1–4.3.4). The purpose of collecting and analyzing process data was to determine a) how automatically generated transcripts affect the work of intralingual subtitlers (represented by user story 4.1.4) and b) how automatic translation of subtitles affects the work of interlingual subtitlers (represented by the combined functionality of user stories 4.3.1–4.3.4).

A process study pilot was carried out in November - December 2019 at the YLE premises. In this study, professional subtitlers created intralingual or interlingual subtitles using ASR output (for intralingual subtitling) or MT output (interlingual subtitling) for selected short video clips. As process metrics, task time and technical effort in the form of keystrokes were compared between two different ASR outputs and two different MT outputs. Subtitlers' subjective evaluations of the usability of ASR and MT for these purposes were also collected using the UEQ survey and semi-structured interviews (see Section 5).

In the interlingual subtitling case the post-editing process was compared to the translators' process when creating the subtitles from scratch. In the intralingual subtitling case the post-editing process was compared both to the subtitlers' process when creating subtitles from scratch and to their process when creating subtitles with the help of offline respeaking. Respeaking was included in the evaluation for the sake of comparison. Some YLE subtitlers use offline respeaking routinely.

Respeaking is a technique to produce subtitles with the help of speech recognition. The respeaker repeats what they hear, adding commands for punctuation, into speech recognition software, which turns the respoken speech into subtitles. Respeaking is widely used for live subtitling, but at YLE it is thus far only used for offline subtitling, where the respeaker/subtitler can correct mistakes made by the software.

8.1 Motivation

The aim of this final evaluation is investigating the efficiency of an automated subtitling process aided by manual correct. We want to learn if such a process chain is more efficient in terms of time and effort compared to a fully manual authoring process. Conversely, considering a fully automated subtitle generation, we want to learn the level of quality of these generated subtitles and whether they can be used for distribution of audiences with minimal or no manual correction. We evaluate these processes for both intra-lingual (supported by an ASR process) and interlingual subtitling (supported by an optional ASR and MT process). In this evaluation, the subtitling test panel used expert subtitle authoring software for the post-editing tasks of the tests.

8.2 User test setup

As with the other two evaluations, 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.

8.2.1 Material used

Suitable video clips to be used as source material for the subtitling experiments were selected from the MeMAD datasets representing two slightly different content types:

1. EU election debates: unscripted dialogue between multiple participants discussing topics relevant to the European elections;
2. Lifestyle or cultural programs: semi-scripted dialogue or in some cases monologue by program host(s) covering various topics (e.g. movies, food and drink, crafts, social issues).

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; and 3) contained approximately 30-35 subtitles. The length and number of clips was limited due to the limited availability of participants for the experiments. Some of the clips selected were short programs already in the three minutes range, while others were cut from longer programs ensuring that they formed a coherent, self-contained section.

Intralingual subtitling

The intralingual subtitling experiments were carried out in Finnish. A total of seven clips were selected from the Finnish dataset: four clips from EU election debates held in 2014, and three clips from a youth-oriented Finnish lifestyle program “Onks noloo”, which contains relatively informal, semi-scripted dialogue between two program hosts.

Transcripts of the original audio were created using two different ASR systems: the baseline ASR included in Limecraft Flow (based on the commercially available Google Speech Recognition¹⁰) prior to the MeMAD project, and the ASR developed by Lingsoft (which is defined in more detail in D2.2). Subtitles for each clip were then generated automatically in Flow, exported into SRT format and imported into the native expert subtitling software used at YLE, for post-editing by participants (cf. Figure 10).

¹⁰ In this case, Google Speech was used instead of Speechmatics, due to the fact that only Google Speech supported Finnish ASR at the time of testing. Cf. <https://cloud.google.com/speech-to-text/docs/reference/rest>; we used v1 of the non-streaming version of this service, calling the *longrunningrecognize* method, with its default speech model.

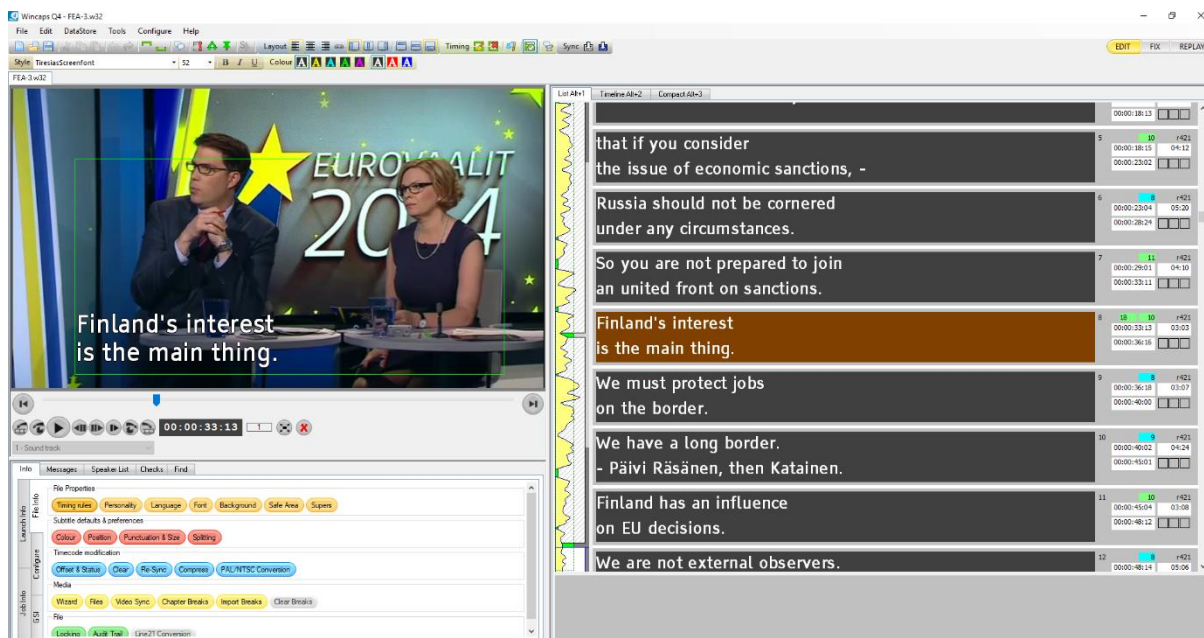


Figure 10: Screenshot of the expert subtitling software used at YLE and for the subtitling evaluation.

Interlingual subtitling

The interlingual subtitling experiments were carried out in four language pairs: Finnish-Swedish, Swedish-Finnish, Finnish-English, and English-Finnish. For the cases with Finnish as source language, six of the same clips selected for the intralingual case were used, three from the EU election debates and three from the youth-oriented program. Similarly, six clips were selected for the other two source languages. For Swedish, these involved three clips from the EU election debates in 2014 and three clips from a lifestyle program called “Strömsö”, which contain semi-scripted monologues by the program host as well as dialogues between the host and a guest. For English, the clips consisted of three clips from European Commission president candidate debates from 2019, and three clips from a cultural program called “Festivaalipuhetta”, which contain semi-scripted monologues by the program host.

For this evaluation round, the machine-translated subtitles were created using human generated intralingual subtitles as the source text. These pre-existing subtitles were translated with two different MT models (sentence-level and document-level model). Subtitles were then generated from the MT outputs using the segmentation and timing from the intralingual subtitles used as source text, and imported into the subtitling software (cf., again Figure 10) in SRT format. For a more detailed description of the MT models and segmentation approach, see D4.2.

8.2.2 Participants

Professional intra- and interlingual subtitlers working with languages in question were recruited as participants. Most of the participants in this round of testing were in-house subtitlers for the project partner YLE, but an additional four freelancers (working outside

of YLE) also participated in the interlingual subtitling experiments. The participation of external freelancers was considered important for broadening the applicability of the evaluation results, and our goal is to include more participants outside of the project partners in the final evaluation phases.

Intralingual subtitling

Four subtitlers participated in the intralingual subtitling experiment. All four subtitlers were in-house subtitlers for the project partner YLE. All participants were professional subtitlers with between 9 and 20 years of experience in subtitling. All four participants had used automatic speech recognition for subtitling before in the form of offline respeaking: three had used it occasionally and one had used it frequently.

Interlingual subtitling

In total, 12 translators (three for each language pair) participated in the interlingual subtitling experiment. In addition to eight in-house translators working for the project partner YLE (Finnish-Swedish, Swedish-Finnish, English-Finnish), four freelancers were recruited as participants: three for the language pair Finnish-English and one for English-Finnish. All participants were professional translators with between 4 and 30 years of professional subtitling experience in the language pair in question. Of the 12 participants, only two indicated they had previously used MT specifically for subtitling, although seven had used MT for other purposes.

8.2.3 User data collection and tasks

The experiments for subtitling data collection were arranged at YLE premises in November and December 2019. The subtitling tasks were carried out using the subtitlers' preferred software environment; most subtitlers worked with the software Wincaps Q4 (cf. Figure 10), two of the freelance translators used Spot software¹¹. To replicate their normal working environment, an external monitor and keyboard were provided, and the subtitlers had access to the internet as well as terminology and other resources normally used in their work.

During the subtitling tasks, process data were logged using Inputlog [7], which records all keyboard and mouse activity. Screen recording software provided as part of Windows 10 was used to capture video of the process to support further analysis. A pre-task questionnaire was used to collect background information from the participants, and post-task questionnaires (cf. Section 5) were used to collect subjective assessment of the ASR and MT output and user experience after each task. After the completion of all 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.

¹¹ Cf. <https://www.spotsoftware.nl/>.

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.

Intralingual subtitling

Three of the four participants carried out seven tasks; the fourth participant did not complete task 4 listed below due to scheduling issues. The following tasks were carried out:

1. Subtitling “from scratch” (2 clips). The participants subtitled the clip without ASR output. Spotting of the subtitles was also done by the participant.
2. Subtitling with respeaking (2 clips). The participants used the Sanelius¹² respeaking software to subtitle the clips, correcting the ASR results as needed. Spotting was also done by the participant.
3. Post-editing of subtitles created automatically with output from the Google Cloud baseline ASR tool integrated by the Flow platform (referred to as “default” ASR in the following text) (2 clips). The participants created subtitles using ASR output and automatic spotting.
4. Post-editing of subtitles created automatically with Lingsoft ASR output (1 clip). The participants created subtitles using ASR output and automatic spotting.

To account for potential differences related to the difficulty of content in each clip, the clips were rotated between tasks in round-robin format. Task order was also rotated to minimize facilitation effect.

Interlingual subtitling

Each participant carried out the following six tasks in the case of interlingual subtitling:

1. Subtitling “from scratch” (2 clips). The participants subtitled the clip without MT output. Spotting of the subtitles was also done by the participant.
2. Post-editing of sentence-level MT output (2 clips). The participants created subtitles using MT output and spotting based on pre-existing intralingual subtitles.
3. Post-editing of document-level MT output (2 clips). The participants created subtitles using MT output and spotting based on pre-existing intralingual subtitles.

To account for potential differences related to difficulty of each clip, the clips and MT outputs were rotated in a round-robin format so that for each clip was subtitled once in each condition (no MT output, sentence-level MT output, document-level MT output) by a different participant. Task order was also varied to minimize facilitation effect.

¹² Cf. <https://www.aanicompany.com/sanelius-ja-kirstu.html>

8.3 Analysis of user data

8.3.1 Productivity data

To assess productivity in the intra- and interlingual subtitling experiments, the process logs recorded during the subtitling tasks were analyzed using Inputlog's analysis functions. Task time and number of keystrokes logged were used as productivity measures. These measures were then compared between the tasks of creating subtitles from scratch and post-editing the ASR or MT output provided. For task time, we observed both total task time and task time subtitling. Total task time reflects the total time taken by the participant to complete an individual task. In addition to creating the subtitles themselves, activity during the task also includes, for example, searching for terminology or other information online. Using Inputlog's analysis functions, we also focused on the time spent specifically in the subtitling software, excluding activity in other windows. For keystrokes, we focused on keystrokes created in the subtitling software, excluding online searches and other activity. In addition to the total number of keystrokes, types of keystrokes were further analyzed based on the process logs to compare effort spent on different activities during subtitling. Text production includes alphanumeric keys related to producing characters, and text deletion includes keystrokes related to removing characters. Keystrokes related to editing the subtitle frames (creating, deleting, splitting or merging frames etc.), as well as changing the timing of subtitle frames, are shown separately. Keystrokes related to other activity such as navigation and video controls are grouped as miscellaneous keystrokes.

Intralingual subtitling

Figure 11 shows the average total task time and average subtitling task time for interlingual subtitling. Task times are shown averaged over the four task types listed in subsection 8.2.3: subtitling from scratch, with respeaking and with two types of ASR output. The comparison of average task times indicates that, on average, 1) subtitling from scratch was the fastest condition, 2) subtitling with respeaking is slightly faster than using ASR output, and 3) no considerable difference in task time was observed between post-editing the two different ASR outputs.

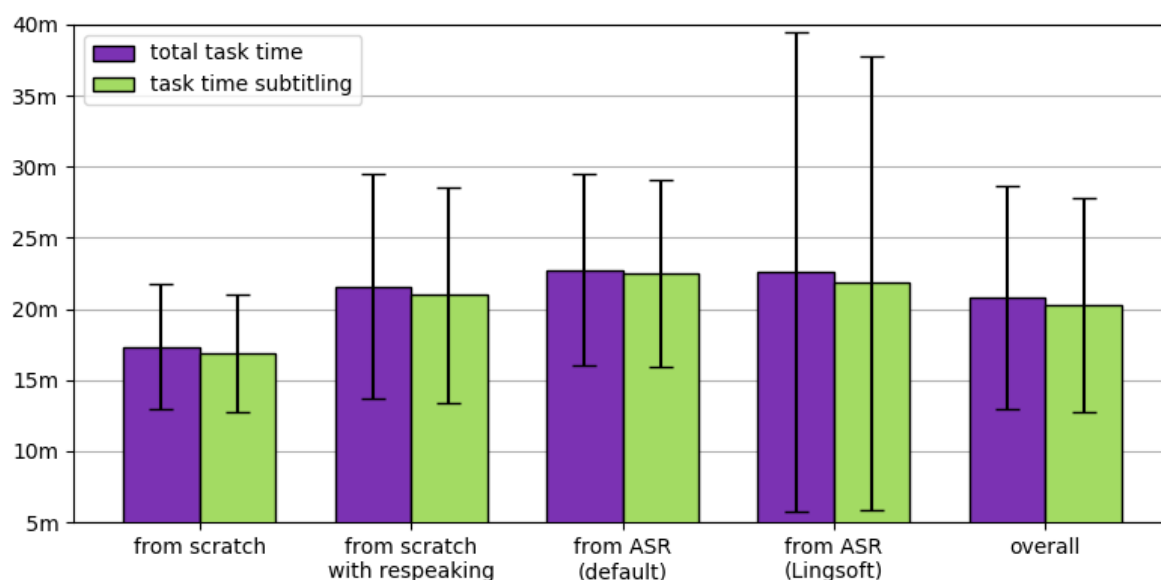


Figure 11: Average total task times (left) and task times subtitling (right) for intralingual subtitling.

Figure 12 shows a comparison of technical effort in terms of the average number of keystrokes used when producing subtitles. The comparison of keystrokes indicates that, on average, 1) producing subtitles from scratch involves a higher number of keystrokes than post-editing ASR output, 2) using respeaking involves the highest number of keystrokes, and 3) post-editing output from the Lingsoft ASR system involves fewer keystrokes than post-editing the basic ASR output. In the case of respeaking, it should be noted that text production includes both keystrokes produced by typing and keystrokes produced through speech recognition. Comparing the types of keystrokes, it can be seen that using ASR output reduces the need for text producing keystrokes, however, the number of keystrokes for text deletion is higher, as participants need to make corrections to the output. The number of miscellaneous keystrokes is higher for the post-editing cases, which may to a large extent be explained by more keystrokes (e.g. arrow keys) used to navigate within the text.

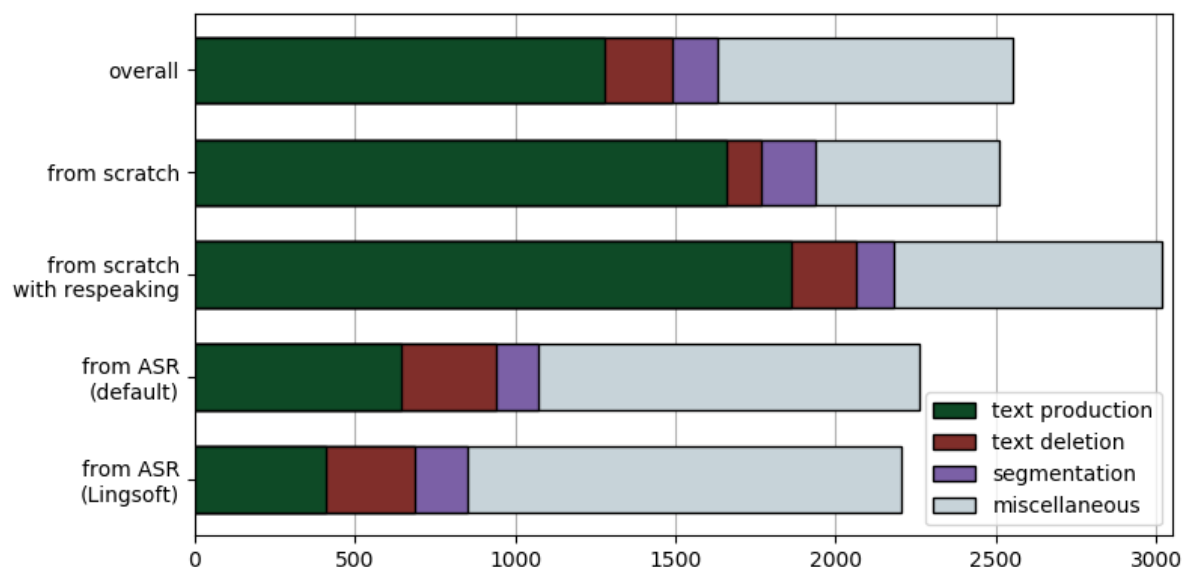


Figure 12: Average numbers of keystrokes, divided into keystrokes for text production (leftmost), text deletion (mid-left), timestamp and segmentation edits (mid-right), and miscellaneous actions (rightmost) for interlingual subtitling.

Variation between the four participants was also observed both in terms of task time and number of keystrokes. Figure 13 shows a comparison of average task times (a) and average number of keystrokes (b) for each participant when they post-edited ASR (left bar) and when they created subtitles from scratch (right bar). An interesting observation can be made that while participants A and B both use fewer keystrokes for post-editing, they are in fact slower when post-editing than when creating subtitles from scratch. Considerable variation can also be observed in the task times and number of keystrokes for individual clips, as shown by the error bars. Due to the variation, definitive conclusions about the effect on task times cannot be drawn yet, but this will be a focus point during subsequent evaluations.

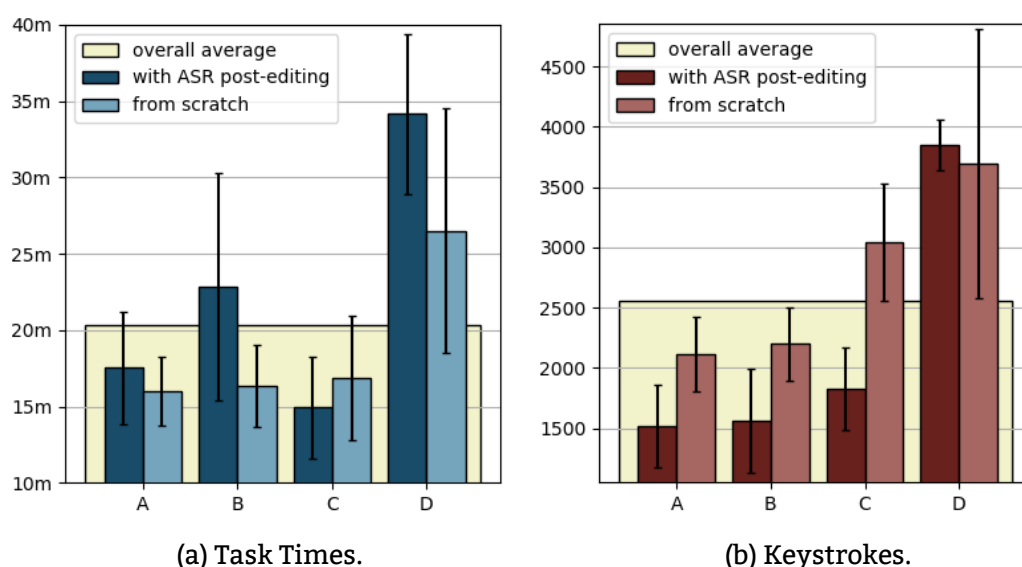


Figure 13: Average task times (a) and average number of keystrokes (b) when post-editing ASR output (left) and when subtitling from scratch (right) for each participant (labelled as A, B, C,D).

To assess the number of changes made to the ASR outputs (default and Lingsoft), edit distance metrics were also calculated. The Word Error Rate (WER) and Letter Error Rate (LER) calculations have been carried out for plain text: First the original SRT files were converted into text files. Sentences that continue from one subtitle block to another and annotated with a hyphen to mark that the sentence continues were automatically joined. Changes in capitalization were counted as errors. WER is derived from Levenshtein edit distance measure, but it operates at the level of words, or rather strings separated by whitespace or punctuation and measures the number of deletions, substitutions and insertions between the ASR result and the subtitle reference. Similarly, LER operates at the letter level.

Table 4 shows the WER and LER scores for each ASR output edited by the participants. The columns list the clip subtitled, the ASR output used and the participant identifier in addition to the two scores. Depending on the clip and participant, the WER scores range from 36.52% to 98.58%, and the LER scores from 13.67% to 51.50%. Edit distance scores are, on average, lower for the Lingsoft output than the default output. While the scores indicate considerable rewriting, it is important to note that when comparing ASR output and final subtitles, not all changes captured by these metrics are necessarily errors in the speech recognition. Subtitles cannot contain everything that is said, but the meaning must be condensed to fit the spatial and temporal limits. Thus, differences calculated by the metrics also contain the amount of necessary condensation for subtitling. Differences are also observed between the participants: participant D appears to overall edit the output more than the others, suggesting variation in individual preferences. A more detailed manual analysis of the changes would be needed to determine the extent to which differences reflect genuine errors in the ASR output and which are due to condensation and related changes made by the subtitler.

Clip	ASR	Participant	WER	LER
EU01	basic	fi-A	58.26%	30.91%
EU02	basic	fi-B	61.03%	41.29%
EU02	basic	fi-D	65.44%	43.00%
EU03	basic	fi-C	47.60%	26.18%
EU04	Lingsoft	fi-A	45.97%	24.35%
EU04	Lingsoft	fi-D	63.13%	39.82%
EU05	Lingsoft	fi-C	36.52%	13.67%
ON01	basic	fi-B	77.56%	40.18%
ON02	basic	fi-A	87.06%	41.09%
ON02	basic	fi-D	86.57%	48.05%
ON03	basic	fi-C	98.58%	51.50%

Table 4: WER and LER edit distance scores between the ASR output and post-edited subtitles for each clip subtitled.

In addition to changes in the text, the participants changed also the segmentation and timing of the subtitle frames. For the basic ASR output, the post-edited files contained on average slightly more subtitle frames (+1%) but fewer lines within the frames (-4%). For the Lingsoft ASR output, the participants reduced both the number of subtitle frames (-6%) and lines (-15%). However, considerable variation was seen between different clips and different participants. When the end and start times of subtitle frames were compared, it was observed that none of the automatically timed frames had been retained in the post-edited versions. For 12% of all automatically created subtitle frames, the participant had accepted either the start or end time, but the majority (88%) of segments had both start and end times differing from the ASR output.

Interlingual subtitling

Figure 14 shows the average total task time and average subtitling task time for interlingual subtitling. The comparison of average task times indicates that, on average, 1) post-editing machine-translated subtitles (regardless of MT output) was faster than

creating subtitles from scratch, and 2) post-editing output by the sentence-level MT system was slightly faster than post-editing the document-level MT output.

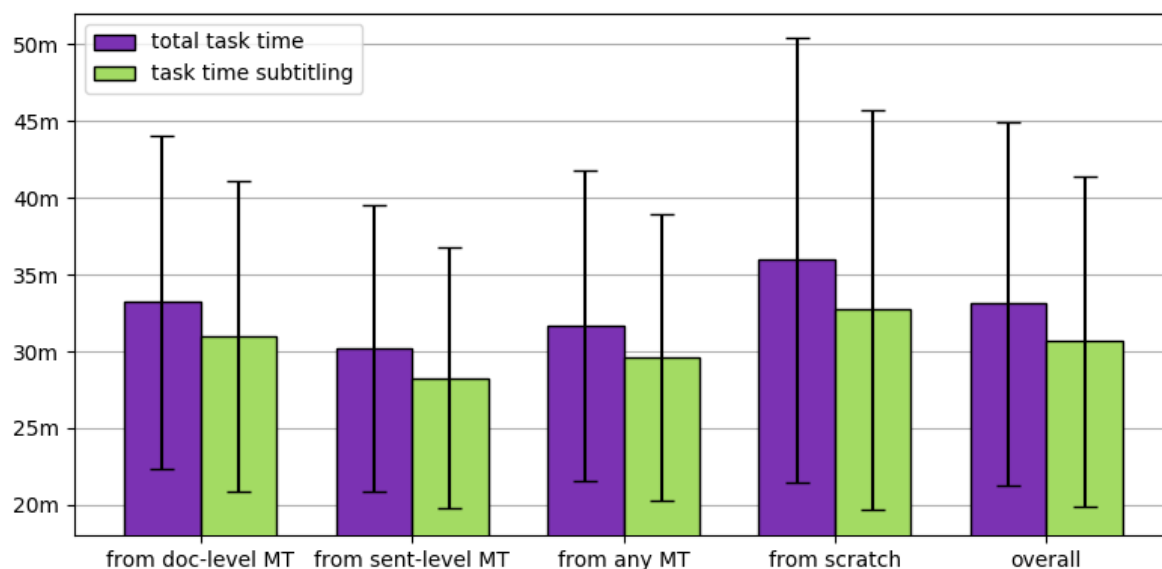


Figure 14: Average total task times (left) and task times subtitling (right) for interlingual subtitling.

Figure 15 shows a comparison of technical effort in terms of the average number of keystrokes used when producing subtitles. Like task times, the comparison of keystrokes indicates that, on average, 1) post-editing machine-translated subtitles (regardless of MT output) involved fewer keystrokes than creating subtitles from scratch, and 2) post-editing the sentence-level MT involved fewer keystrokes than post-editing the document-level MT output. Some differences can also be seen in the distribution of keystroke types. Intuitively, post-editing reduces the need for text producing keystrokes compared to writing the subtitles from scratch. However, the share of text deleting keystrokes is somewhat higher, as correcting the output also involves removing words or characters. The number of keystrokes related to editing subtitle frames can also be expected to be higher in the from scratch mode, as the participants needed to create and set the timing for each subtitle segment themselves.

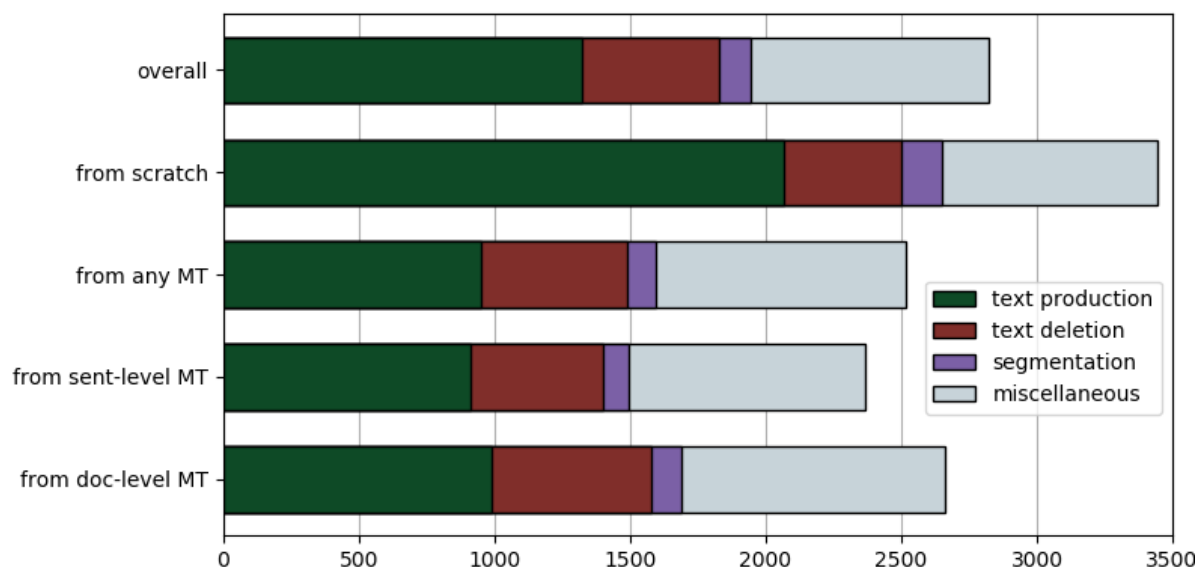


Figure 15: Average numbers of keystrokes, divided into keystrokes for text production (leftmost), text deletion (mid-left), timestamp and segmentation edits (mid-right), and miscellaneous actions (rightmost) for interlingual subtitling.

Observations regarding the two different MT outputs, changes to subtitle frame segmentation and variation between different post-editors are discussed in more detail in Deliverable D4.2.

8.3.2 User Experience Questionnaire

Subjective evaluations regarding the use of ASR or MT output and the post-editing experience were collected after each post-editing task using the adapted UEQ questionnaire (see Section 5.1). For the purposes of the subtitling cases, the questionnaire focused on post-editing and correcting the spotting of subtitle frames.

Intralingual subtitling

Figure 16 shows the UEQ scores for each question averaged over all participants and all clips in the two post-editing tasks (default ASR output and Lingsoft ASR output) as well as in the respeaking task. The value 0 represents a neutral mid-point between negative and positive evaluations using the adjective pairs (e.g. “Post-editing was unpleasant” vs “Post-editing was pleasant”). Based on these evaluations, the participants appear to generally prefer respeaking over correcting either of the two ASR outputs. Comparing the two ASR outputs, the Lingsoft ASR output is generally evaluated in more positive terms, except that the participants found it more limiting than the basic ASR. While the participants indicated that editing the ASR output, particularly the Lingsoft ASR, was relatively easy, fast, simple and to some extent efficient, they also characterize the experience as relatively boring and limiting. In the questions regarding automatic spotting and segmentation, the participants appear to evaluate these aspects are relatively poor but easy to correct.

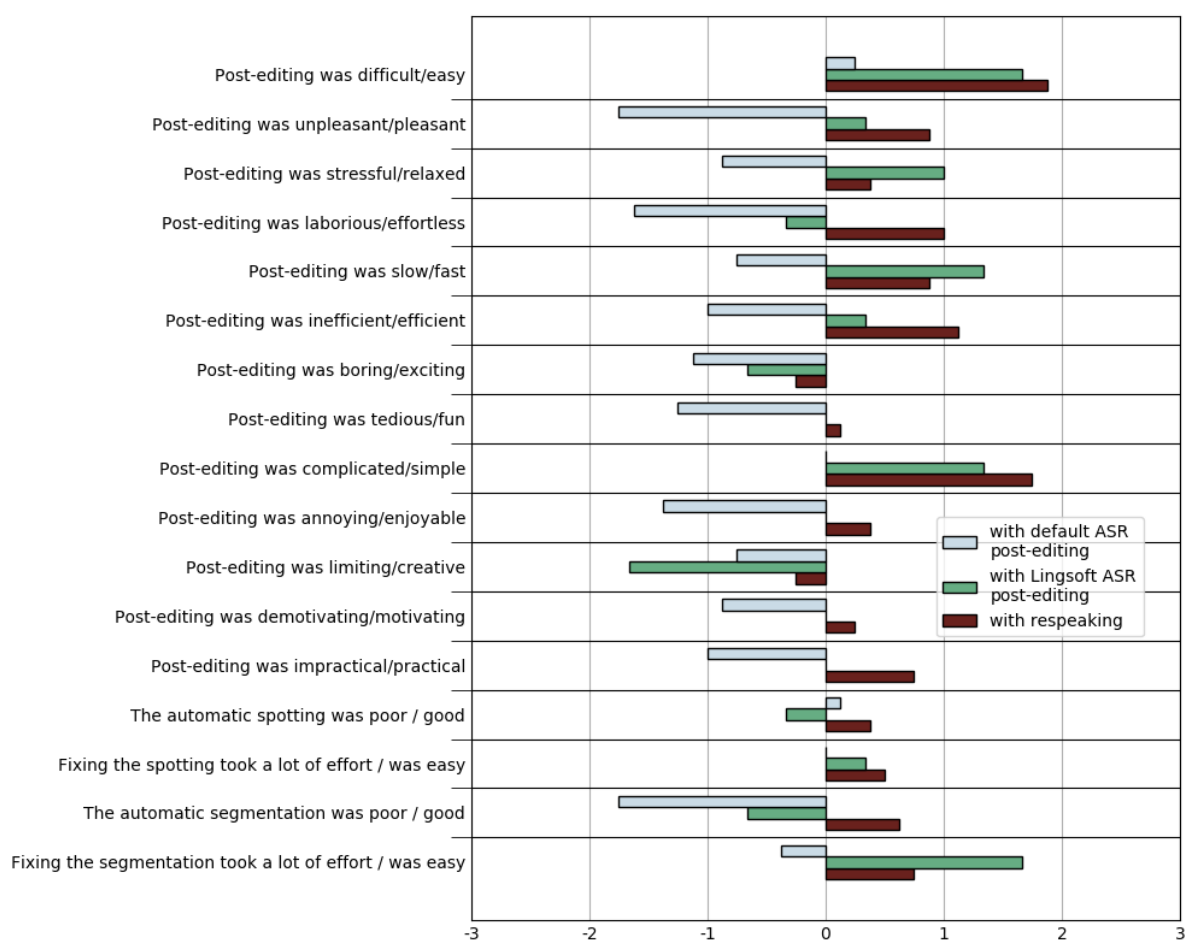


Figure 16: Average UEQ scores for intralingual subtitling, comparing post-editing of basic ASR output (top, light blue), post-editing of Lingsoft ASR output (middle, green), and respeaking (bottom, red).

Interlingual subtitling

Figure 17 shows the UEQ scores for each question averaged over all participants and all clips comparing the two MT outputs (sentence-level model and document-level model). The value 0 represents a neutral mid-point between negative and positive evaluations using the adjective pairs (e.g. “Post-editing was unpleasant” vs. “Post-editing was pleasant”). Based on these evaluations, the participants appear to consider the MT post-editing experience overall rather neutral or negative. While they do not appear to consider post-editing particularly difficult or complicated, there is a slight tendency to describe it as, for example, somewhat laborious, annoying and limiting. There appears to be a slight overall preference for the sentence-level system, but the differences are very small. The most negative responses can be seen in the questions relating to spotting and segmentation of the subtitle frames, which the participants found poorer and more difficult to correct in the case of the document level model.

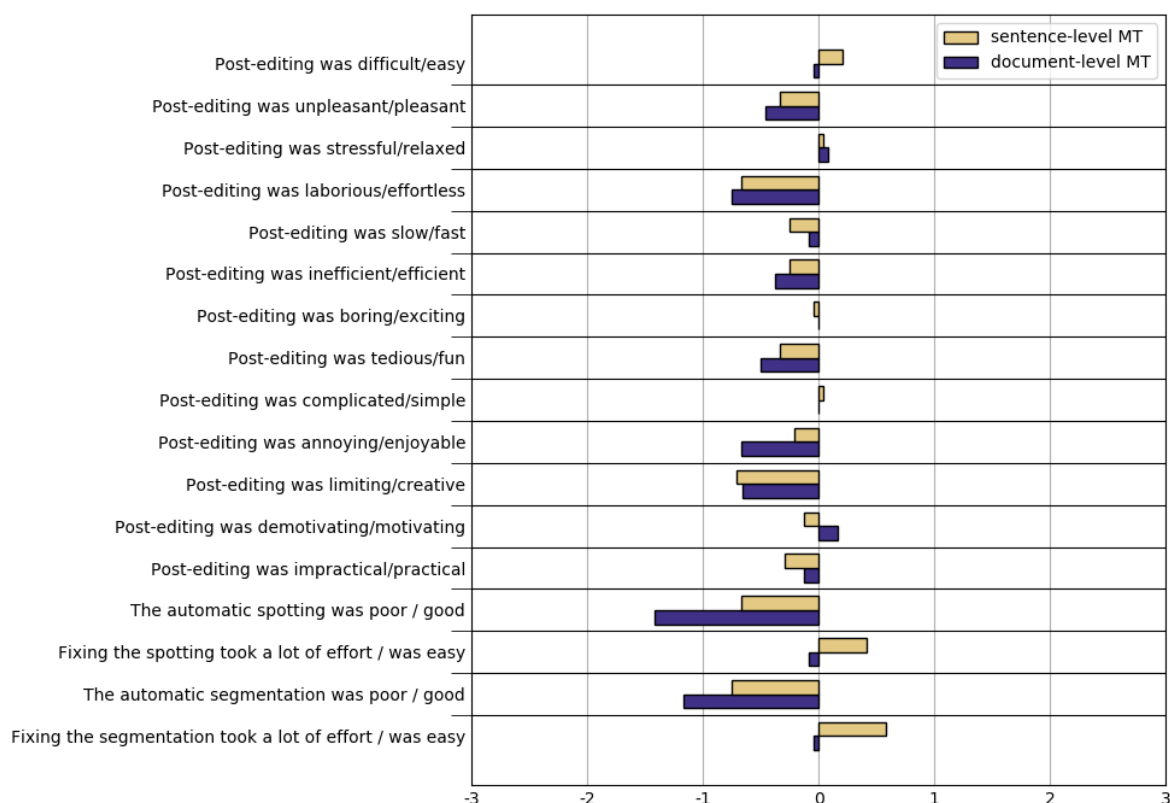


Figure 17: Average UEQ scores for interlingual subtitling, comparing post-editing of sentence-level MT (top, yellow) and document-level MT (bottom, purple).

Interestingly, participants post-editing different language pairs appear to differ in their experiences. Figure 18 shows average UEQ scores for MT post-editing separately for the four language pairs English-Finnish, Swedish-Finnish, Finnish-English, and Finnish-Swedish. On average, evaluations by the translators working with English-Finnish or Swedish-Finnish appear to be more negative than in the two language pairs where Finnish was the source language. Translators working with Finnish-English MT output characterize post-editing in neutral to slightly positive terms as easy, relaxed, efficient as well as exciting and motivating. Particularly negative responses can be seen regarding the automatic spotting and segmentation, which was found poor and difficult to correct especially in the English-Finnish case. This observation may relate to specific issues identified in the segmentation and timing changes, which are discussed in more detail in Deliverable D4.2.

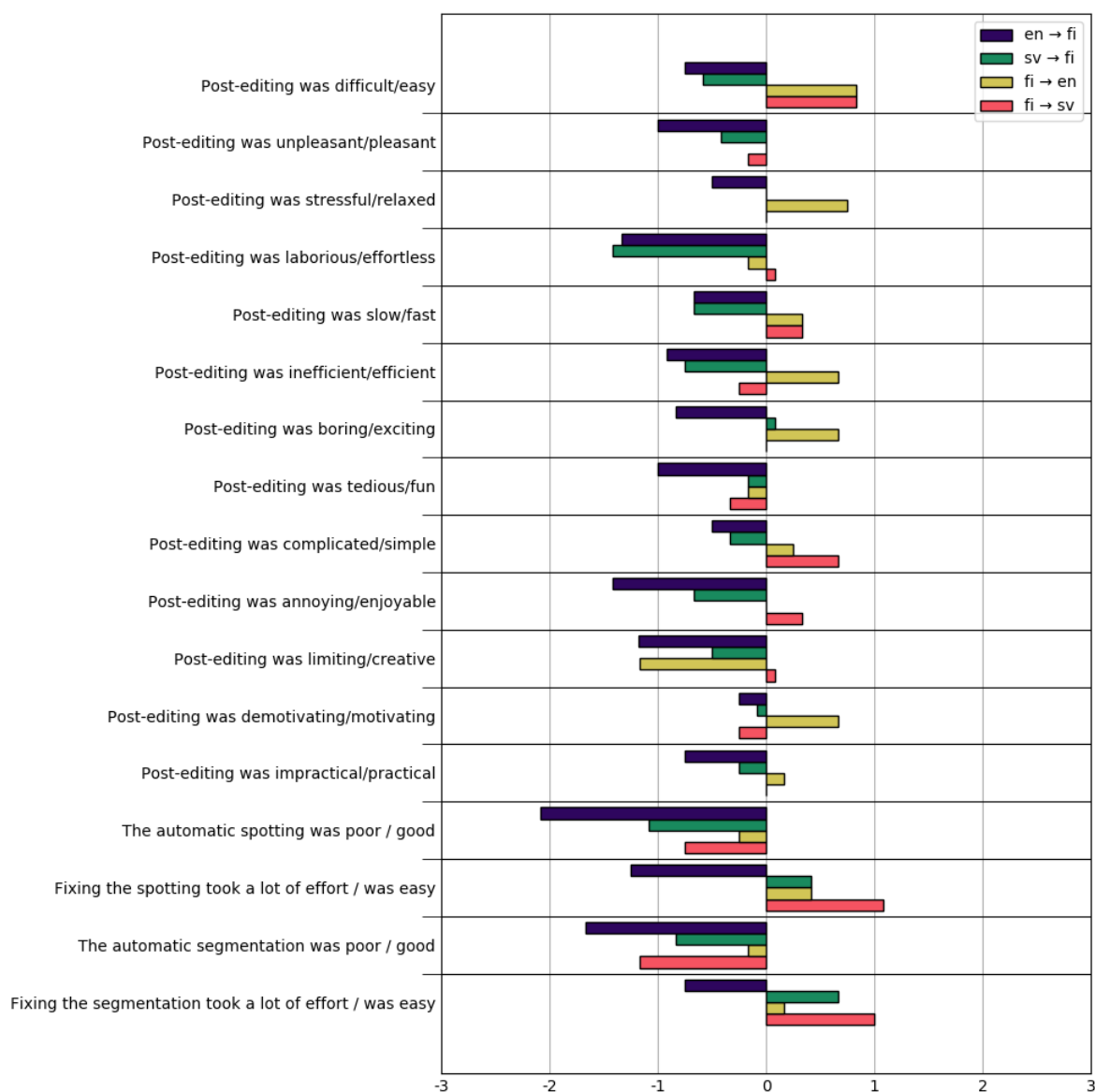


Figure 18: Average UEQ scores for interlingual subtitling, comparing English-Finnish (top), Swedish-Finnish (second from top), Finnish-English (third from top), Finnish-Swedish (bottom).

8.3.3 Feedback from interviews

For the subtitling functionality evaluation, the following outline was used for the semi-structured interviews:

1. How did the post-editing tasks feel overall?
2. Was there anything positive/negative about the tasks?
(Depending on question 1: If the participant's overall answer was negative, ask about positives and vice versa.)
3. What features of the ASR/MT output impacted the post-editing most, in good or bad?
4. Did you notice any differences between the ASR/MT outputs?

5. What is your subtitling process usually like with short clips like these?
6. How did the use of ASR/MT output impact your own work process?
7. Could you imagine using ASR/MT as a tool for subtitling?
8. How should the ASR/MT be improved as a tool?

Intralingual subtitling

The interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements. A total of 30 negative comments, 12 positive comments and 8 mixed comments were identified in the interview transcripts of the four participants. Most comments were not linked to a specific system output, instead they referred to the outputs or the post-editing experience in general. Most identifiable comments referred to the output of the software used for respeaking (5 positive, 5 negative, 1 mixed) or the “default” ASR output (8 negative, 2 mixed). Comments regarding the Lingsoft ASR output were fewer in number (1 negative, 2 positive), which is likely explained by the fact that it was used for only one task which was completed by three participants.

Of the negative comments, half (16 statements) referred to overall speech recognition errors. Specific issues mentioned by the participants involved incorrect spotting and timing of subtitles (3 statements), failure to recognize proper names (2 statements), as well as errors in compound words, punctuation, recognition of similar words and omitted words. Negative comments regarding the effect of using ASR output (13 in total) related to post-editing causing more effort (8 statements), being more difficult (2 statements), leading to more need to navigate around in the text, as well as perceptions of the ASR output as frustrating and limiting. Of the negative comments, 4 identified the more colloquial language use in the youth-oriented program to cause problems for the ASR.

Most positive comments related to the effect of using the ASR output, characterizing it as easier (3 statements), reducing the need for typing (2 statements), in general having a positive effect on the participant's own process (3 statements) and being interesting. As specific issues related to the output quality, observed omissions of repetition or “unnecessary” words were mentioned as a positive feature (2 statements), handling of proper names and compounds, as well as overall fewer errors than expected. Mixed statements noted that although sometimes the ASR output reduced the need for typing or made the process easier, at other times output quality was so poor it led to more effort (2 statements) or frustration.

Two of the four participants indicated they would use ASR and post-editing for intralingual subtitling, stating that they found it easier. One would not consider using ASR for this purpose due to the quality currently being too poor, and one gave a more mixed answer indicating that ASR and post-editing could be suitable for some content, but not for colloquial speech. As suggestions for future development and improvement, the participants mentioned overall speech recognition quality, but focused mainly on segmenting the output into subtitles. They expressed wishes for better sentence segmentation (2 statements) or segmentation based on speaker changes (2 statements),

although one participant also mentioned that having the ASR output separately without segmentation would be more helpful.

Interlingual subtitling

The interviews were analyzed for positive and negative statements, specific issues raised by the participants, and potential suggestions for future development and improvements. A total of 79 negative statements, 42 positive statements and 22 mixed statements were identified in the interview transcripts. No clear differences were observed between the two different MT outputs. In most cases, the statement concerned the MT outputs in general and no specific output could be identified. In the cases where a specific output was identified, roughly equal numbers of statements were deemed positive (document-level 7, sentence-level 6) negative (document-level 13, sentence-level 14), or mixed (document-level 5, sentence-level 7).

Of the negative statements, the majority (33 statements) concerned the spotting or segmentation of subtitles, which was identified as a problem by all 12 participants. Specific issues mentioned by the participants concerned translations being out of sync with the audio, and cases where a sentence had been incorrectly split into two (or more) segments. Two participants also brought up a related issue of condensation, noting that the MT output tried to pack "too much" into a given subtitle frame and that the machine is not able to condense the translation. Although the translations were created based on intralingual subtitles, which often already involve some condensation compared to the audio, this may suggest further differences between different languages. On the other hand, two participants also noted omissions (missing words or longer passages) in the MT output, although it was not entirely clear whether this was caused by information being omitted by the intralingual subtitler or missing words in the MT output. MT output quality received 19 negative mentions. Specific issues mentioned by the participants involved lexical errors (mistranslated or "odd" word choices) and accuracy errors involving the meaning of longer passages, as well as fluency issues (ungrammatical or unidiomatic structures). In 11 cases, the statement referred to the output but it was not possible to determine whether the comment concerned the linguistic quality of the MT or the segmentation. With regard to the effect of using MT output on the participant's own processes, negative statements characterized the effect as reducing productivity or being more effort than translation from scratch (15 statements), limiting the participant's own processes (8 statements) and potentially leading to lower quality of the final translation (5 statements), or characterized the experience in general negatively without a clearly specified issue.

Positive statements involved general positive characterizations of the output without defining specific issue (11 statements). Most positive statements defining a specific issue concerned lexical issues (9 statements), with the participants commenting that they were positively surprised by how much useful terminology they were able to get from the MT output. Fluency of the output was mentioned positively in 3 cases. Two positive statements were made regarding the subtitle segmentation, in both cases mentioning that in some outputs it was "better than in the other one". The remaining positive statements involved either general positive characterization of the output as good or useful without specifying issues related to linguistic features or spotting (11 statements)

or comments characterizing the overall experience in positive terms (12 statements). Further 22 statements were identified as mixed, in that they were not clearly negative or positive. These statements were mostly generic, referring to undefined aspects of the MT output or the experience overall. A typical mixed statement was made by participant f1svA, who noted that the output was "sometimes surprisingly good but sometimes surprisingly bad".

Of the 12 participants, 4 indicated they would consider using MT as a tool for subtitling, although all stated they would like to see some improvements in quality. Two participants stated they could not see themselves using MT as a tool, while 6 gave a more mixed answer, stating that they would consider MT and post-editing suitable for some situations but not others, for example, depending on the type of program and subject matter. Comments regarding the situations where they considered MT usable varied. Some commented they would see it most useful for content they were not very familiar with, for example, to assist with terminology, whereas others stated they would only use MT with subject matter they were already familiar with, to make sure they would spot possible errors. With regard to genre, some stated MT seemed more useful for the EU genre, with more formal speech, while others noted it would be more suitable for "simpler", less formal style.

Regarding future development and improvements, 28 statements were identified in the transcripts. The most commonly mentioned improvement needed concerned spotting/segmentation of the subtitles (8 statements). A further two statements noted that segmentation according to speaker changes would be useful. Four statements referred to making the system work more like a translation memory, so that individual translators or a group of translators could add their own material and see how things were translated previously, and 4 statements suggested integrating some type of terminology resources into the system. In two cases, participants commented they would rather see the MT output separately as a whole without segmentation, and 1 mentioned showing ASR output of the original audio. Other specific issues mentioned involved need for condensing the MT output for subtitles (2 statements), improving cohesion, genre adaptation, general presentation of the subtitles, and overall quality. Two of the participants also specifically mentioned the multimodal aspect of the situation, one remarking that the machine is not able to take visual information into account and the other wondering whether it would be possible that the MT uses visual information.

9 Discussion and impact of the second prototype evaluation

In this section, we summarize the overall results of the second end-user evaluation round in the project. First, we provide a general conclusion for each evaluated set of functional user stories. We then explain how the observations made of the evaluation sessions and direct feedback gathered from the test panels will be incorporated into the project's execution in the following ways:

1. The results impact the metadata use and formats specifications, which we will define in their final form in D6.7, ensuring that all features and requested data flexibility are correctly represented. On the other hand, this feedback will no longer impact the selected use cases (this selection was made at an earlier stage in the project but will also be reported in D6.7).
2. The results will impact the implementation and development of the final iteration of the final integrated prototype's interfaces and functionality. Based upon given feedback, we will improve aspects of the prototype that underperformed or were still missing in the second prototype.
3. Finally, there's impact on the final round of evaluation of the prototype, which has become clear from executing these first batch of user tests. Tweaking the evaluation procedures for 2020 will be especially important given the broader test panel audience, as usage of the platform will be shared between WP6 and WP7 for its dissemination activities.

We discuss each impact topic in each of the following subsections.

9.1 Overall observations of the second evaluation round

Before we discuss the details of the observations and feedback gathered during our evaluations, we present the main take-aways for each functional epic.

9.1.1 Overall conclusions concerning the evaluation of "Editing assistance using multi-modal and multi-lingual metadata"

Participants were enthusiastic about the new technologies being evaluated and found the available metadata useful, but at the same time they did not find it very relevant for their daily work. Normally video editors at YLE receive a detailed script to work with from the journalist, with exact time codes for the segments needed. When such detailed scripts are provided, the video editors don't need to search for the content, as they can just go to the right part of the material directly.

Regardless, two out of three participants stated they could imagine themselves using such metadata in their work, even if cases where it would be needed might be few. As suggested by the participants, the kind of metadata evaluated in this case might be more useful to the journalists writing the scripts, or perhaps to journalists who edit their own videos, than to the video editors. Further evaluations should shift focus in that direction, to ensure that we can test the use of this generated metadata to its fullest before making

conclusions about any efficiency increases that can be realized in the video editing process.

Quick improvements can be made to generated metadata to make them more relevant though. This includes combining multiple modalities (e.g., speech transcripts with recognized faces) to make finding relevant sections easier – or even possible at all – using the limited search capabilities of the editor’s craft editing software. Reducing the amount of metadata using summarization would also aid in helping users to keep oversight in the wealth of low-level metadata as well as allow the editing software to remain more responsive.

Regardless of the assessment of descriptive metadata usability, we will do more evaluations specifically to gauge the usefulness of machine-translated transcripts to aid in the video editing process when editors need to assemble foreign-language content.

9.1.2 Overall conclusions concerning the evaluation of “Searching and browsing for ingested and archived content”

The participants were somewhat frustrated with the tasks, finding some of them difficult, but much of their feedback was positive. In this evaluation the participants used the Flow platform, which they were not previously familiar with, and their unfamiliarity with the platform and its search logic affected the results. The participants also noted that a search could result in too many hits, “too much data”, and some kind of filtering was needed. Even though many search capabilities are already available, it was insufficiently obvious to the candidates how to efficiently use these features. This is something we need to correct in future evaluations.

Search results would also have been much improved if the participants had had the option to search based on content descriptions or images instead of just words and phrases in the transcripts. Despite the criticism, all participants stated that they would use this kind of metadata in their work, some of them quite enthusiastically. Work on incorporating content descriptions in the content retrieval functionality will continue in 2020.

9.1.3 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 panel did not clearly grade the process as being more efficient, or as 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. It remains to be evaluated in which scenario the most gains can be made: whether this will be in a professional post-editing workflow, or for the complete automation of a subtitling

workflow with limited manual corrections if the quality is deemed sufficient by consumer test panels.

For interlingual subtitle post-editing, 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.

9.2 Impact on metadata usage and format specifications

We summarize the specific impact the second evaluation round will have on various aspects of the MeMAD project, the first of which is the use of metadata and interchange format specifications. The impact in this case is limited as no obvious shortcomings or extensions have been identified that were not yet described by D6.4. The table below summarizes our findings and subsequent impact.

Evaluation	Observation	Impact
Epic 6.2 and 6.5 (Searching + Editing)	<p>A desire for additional visual description of the edited and search from the generated metadata was noted from many of the test subjects.</p> <p>Similarly, in many cases, the nuisance of too much data to search through was raised by the testers.</p>	Work on T6.1 and T6.2 continues to finalize the processing requirements and interchange formats for visual content descriptors. These will serve as the basis for the final project year's implementation of the captioning, face recognition and summarization components for the MeMAD platform. These results will be reported in D6.7 and the final format specifications will be published on the MeMAD Git repository.

Table 5: Second evaluation round impact on metadata usage and interchange format specifications.

9.3 Impact on the future evaluation of the prototype

This second evaluation round, which for the first time was constructed around the MeMAD prototype and the various algorithms it integrates, has also learned the consortium partners a lot on how to improve the evaluation process itself or has given a clear indication on which additional evaluations should take place to fill in gaps observed in this second evaluation. We list the impact on the final project's year evaluations in the table below.

Overall, we recruited 3-5 participants in the usability study with editing and searching tasks. At this stage in development, the number of participants is enough because research has found that 4-5 users representing one audience segment is enough to reveal

about 80 percent of the most significant usability problems observed by that audience [8]. To avoid as much bias as possible, a wider audience will be recruited overall for the final year's evaluations, including by the dissemination efforts in WP7 as proofs-of-concept will join the formal evaluations in a source of end user feedback.

Evaluation	Observation	Impact
Epic 6.2 (Searching)	Considering that the evaluation panel used the Flow interface exclusively during this evaluation, we observed that the users' unfamiliarity with the GUI did prohibit some participants to fluently execute the given tasks.	While we initially estimated the GUI to be sufficiently intuitive for the given search tasks, and those tasks were also trailed by members of the consortium to estimate their feasibility, the unfamiliarity of users with the Flow platform did hinder the fluent execution of all tasks for some participants. We will take time during future evaluations to more extensively introduce the participants to the interface capabilities and search features of the platform, a.o., through an extensive demonstration, to ensure that this situation will be less an issue.
Epic 6.2 (Searching)	The participants unfamiliarity was not only with the Flow interface itself but also with the metadata stored within the system. The metadata originate from other sources than the manual archiving typically employed and as such they are different in structure and abundance. In particular, the participants noted that a search could result in too many hits, "too much data", and better filtering was needed. Search results would also have been much improved if the participants had had the option to search based on content descriptions or images instead of just words and phrases in the transcripts. Despite the criticism, all participants stated that they would use this kind of metadata in their work, some of them quite enthusiastically. Work on incorporating content descriptions in the search will continue in 2020.	Both visual content descriptions and content summarizations will be incorporated into the final version of the prototype platform. We will take advantage of this new functionality to clearly demonstrate evaluation panels how to employ this functionality (cf. also the previous point) such that the issues of "too much data" and the need for visual content descriptions can be mitigated.
Epic 6.2 (Searching)	Due to the high perceived complexity of the content retrieval tasks, there was little	We will include a future evaluation with more attention question of determining the extent that

	time to perform a clear effectiveness test between the legacy and auto-generated metadata.	metadata auto-generated within the MeMAD prototype can replace human-curated archivist metadata.
Epic 6.11 (Subtitling)	<p>Testing subtitling of different genres of audiovisual material was useful, and suggests that ASR and MT post-editing may be more feasible for some types of content than others.</p> <p>In particular, more colloquial style of speech, such as in the youth-oriented program clips used for the Finnish-English and Finnish-Swedish language pairs may be more challenging for automatic processing. Particularly with the interlingual subtitling case it is important to note that the participants did not have prior experience with using MT for this purpose. This is due to MT post-editing being, in general, still relatively infrequent in Finland, and particularly for audiovisual translation. To some extent, the results from the post-editing experience may therefore reflect unfamiliarity with the task, which some of the participants also commented on themselves.</p>	<p>This observation impacts the future evaluation in the following ways:</p> <ul style="list-style-type: none"> - Further evaluations are therefore being planned for the year 2020 with longer programmes being processed in more realistic scenarios. We are considering the possibility of more long-term pilot testing of the Lingsoft ASR system in combination with Limecraft's subtitle generation and editing in real production environments such as at YLE. • Future evaluations have been scheduled with more extensive panels, sourced a.o., from the ECG member organizations to avoid any Finnish bias in the test results and to test a wider variety of language pairs to evaluate differences in error rates between languages (cf. Section 10.1).
Epic 6.11 (Subtitling)	In this round of subtitling evaluations, interlingual subtitling was always started from existing same-language source subtitles. This is a valid scenario, but it still requires the source subtitles to be available. As such, the outcome of the evaluation doesn't show us yet how manual effort is best spent: in correcting source transcripts, in authoring source same-language subtitles, or in post-editing the final result of a chain of automated generation processes.	As part of the 2020 evaluations, we will deploy the entire suite of metadata editing tools available now in the MeMAD platform, including ASR transcript and subtitle editing (and editing in both the source language, or editing a machine translation). As such, we will be able to determine the optimal place for manually post-editing various metadata, and when to introduce which translation and subtitle generation step. It could be beneficial to first fix the transcript, after which the subtitle generation and automated spotting could perform better. This approach then needs to be compared to an editing environment in which the test panel has more experience, as the GUI used will be different in the Flow platform.

Epic 6.11 (Subtitling)	The current subtitling evaluation was carried out by professional subtitlers who applied their authoring (or post-editing) skills in a high-quality approach with the aim of producing broadcast-ready high-quality subtitles. The question remains at this point what subtitle quality would be good enough in order to be comprehensible and sufficiently readable by consumer (viewer) end users.	We will organize evaluation sessions with consumer end users to gauge the subjective quality of subtitles generated in a variety of ways including different types of manual corrections (cf. Section 10.1).
Epic 6.5 (Editing)	The evaluation participants were enthusiastic about the new technologies being evaluated and found the available metadata useful, but at the same time they did not find it very relevant for their daily work. Normally, video editors at YLE receive a detailed script to work with from the journalist, with exact time codes for the segments needed. As long as such detailed scripts are provided, the video editors don't need to search for the content, as they can just go to the right part of the material directly.	<p>As suggested by the participants, the kind of metadata evaluated in this case might be more useful to the journalists writing the scripts, or perhaps to journalists who edit their own videos, than to the video editors. Further evaluations should shift focus in that direction, as follows:</p> <ul style="list-style-type: none"> • We will schedule further evaluations with journalists rather than with pure video editors to attempt to make the tasks more meaningful, and then re-evaluate if a significant interaction with content metadata effectively takes place in editing. If it doesn't, then we will shift focus for the video editing epic to the interlingual video editing use case. • We will schedule further evaluations with a stronger focus on user story 2.1.6 ("Use of autotranslated content for editing") with members of the ECG (in particular, Hotel Hungaria, TV2 and RTS) for whom the presence of Finnish and Swedish will be a bigger challenge than with the YLE evaluation panel if they don't speak either of both languages.

Epic 6.11 (Subtitling)	It is encouraging to observe that there is already a time gain in the authoring of interlingual subtitles vs. from scratch authoring. However, this evaluation started with intra-language subtitles that were already available. We need to investigate if the same time gain can be repeated when starting from ASR-made transcripts. On the other hand, more time can potentially be gained when machine-translating subtitles if the timing issues encountered can be resolved.	Additional interlingual subtitling evaluations will be done starting from ASR transcripts and with the timing and spotting algorithms available from the Flow platform to re-assess interlingual subtitling generation productivity.
---------------------------	---	--

Table 6: Second evaluation round impact on the future evaluation of the MeMAD integrated prototype.

9.4 Impact on the development of the final MeMAD integrated prototype

In this final impact session, we discuss how the second set of evaluations will crucially impact the final version of the MeMAD prototype and the various algorithms it integrates. Based on the given user feedback and observations of the execution of various evaluation tasks, we can formulate desired improvements to the components being developed by the consortium. We will also use the think-aloud verbalizations and screen captures recorded during the evaluations for further guidance in defining the exact modifications required to the MeMAD prototype's user interfaces and flows of system/user interaction.

We list the impact on the final project's year software developments in the table below.

Evaluation	Observation	Impact
Epic 6.11 (Subtitling)	We observed that many corrections were made to the timecodes of auto-generated subtitles (88% of all subtitles had timing changes applied to them) before the subtitling panel considered them completed and of sufficient quality.	We will further investigate relevancy of timecode changes between initially suggested subtitles and post-edited subtitles. A more in-depth study of these manipulations will help us optimize the currently used subtitle spotting algorithm to divide and time transcription text into subtitles.
Epic 6.11 (Subtitling)	For intralingual subtitling, ASR with post-editing shows promise as a workflow, with most participants indicating they would be interested in using it further. However, they also wished for better sentence segmentation or segmentation based on speaker changes.	As part of the integration efforts in 2020, we will incorporate new features available from the Lingsoft ASR, including speaker diarization for representing individual speaker turns in speech transcripts. This functionality will help the subtitle generation process on key issues raised by the

	Finally, there were complaints that proper names or compound words were often not recognized correctly.	evaluation panel. Additionally, we will also consider whether the Lingsoft ASR can be extended with support for domain-specific dictionaries such that proper names (that could be very genre- or program-specific) can be input into the ASR process on a per-execution basis.
Epic 6.11 (Subtitling)	From the evaluation results and user feedback, we conclude that using pre-existing intralingual subtitles as the source text for MT appears is feasible approach, although further improvements in quality and usability are needed. In particular, the transformation and translation process of source subtitles performed by University of Helsinki did not yet include the domain-optimized spotting algorithms developed by Limecraft in this evaluation. As such, the MT output becoming de-synchronised with the audio became the specific issue most frequently commented on negatively, and both the process metrics and the subjective UEQ assessments indicate that considerable effort was needed to correct the subtitle segmentation and timing.	It is obvious that subtitlers should not spend considerable effort on skewed timing corrections, which is an issue that can be corrected. Subsequent evaluations will incorporate the spotting algorithms built and optimized as part of Limecraft's tools to incorporate better timing alignment with audio for translated subtitles.
Epic 6.11 (Subtitling)	As part of the interlingual subtitling evaluation, the question was raised if the incorporation of a feedback loop was considered, e.g., to optimize translations within a context, e.g., a group of translators working for a given program.	Participants from WP6 and WP4 will investigate whether a feedback loop can be built in the remained of the project, and if the expected gain from this functionality will outweigh its potential disadvantages (incl. complexities in training updated models, ensuring that clean corrections are added to the models without disturbing translation performance, etc.).
Epic (6.5) Editing	Editors made the remark that it would be more valuable to have metadata modalities linked up such that searches can yield better results, especially if the search interfaces they need to use are more limited (which is certainly the case for Avid software). E.g., combining speech	While not strictly correct in many cases, we will investigate combinations of metadata occurring at the same time on a content clip's timeline. As such, we will combine modality that are not explicitly linked together as single metadata that will aid in a more fluent editing process. Some of the

	transcripts with the speaker name from face recognition in a single metadata element can help better find relevant parts of the media.	metadata fusions will be incorrect (e.g., a person shown on screen is not necessarily the one speaking at that time) but evaluations will show if the gains outweigh the disadvantages in this case.
Epic (6.5) Editing	Clear problems were observed with the amount of metadata that was imported into Avid Media Composer and the way this metadata was presented to the users. The limited metadata view and extent of data made manipulation of the metadata (search, viewing) cumbersome and from time to time the software actually worked quite unstable (or even crashed).	As part of the 2020 evaluations, we will experiment with putting the same set of metadata in other relevant editing software such as Adobe Premiere. While this is not the most commonly used video editing software, it is still a craft editing tool used by professionals, and has more potential in terms of metadata support and integration possibilities. A more elegant display and handling of metadata could help improve on the functionality provided for implementing these user stories.
Epic 6.2 (Searching)	We observed that many participants tended to search archives by means of categories or topics assigned to content from controlled lists of vocabularies. For example, when looking for content related to wind energy, the primary search term would have been “renewable energy”, which was not present in the auto-generated metadata using the currently available ‘literal’ metadata extraction methods.	As archivists have the tendency to both generalize metadata terms they enter into an archive using controlled vocabularies and also summarize content into sizable chunks, this presents a clear argument for the adoption of content summarization (as introduced as a new user story 2.2.6 in D6.4). This would allow transliterations of literal terms obtained from aural and visual descriptors into more commonly used topics. As such, auto-generated metadata added to the archive would be better aligned with how archive users currently browse existing archives. Topic detection and summarization will be a main topic of research and implementation for WP3 and WP6 in the final project year.
Epic 6.2 (Searching)	Participants were somewhat frustrated with the tasks, finding some of them difficult. In part, a cause identified for this was the use of the Flow platform which the participants were not previously familiar with, and their unfamiliarity with the platform and its search logic affected the results. This also holds for the non-obvious	Like we discussed in subsection 9.3, in addition to a more extensive training for the use of the search interface, we will capture specifically feedback from the test panels on where and how the ‘search’ user interface could be improved to work more intuitively. Additionally, the incorporation of summarization techniques can

	application of the filtering capabilities already offered by the platform's search interface, which could help in filter away irrelevant metadata from search results.	hopefully also aid in reducing redundant or irrelevant search results, such that only relevant topics are retained and the presented metadata is less granular and massive to begin an initial triage of search results.
	Participants indicated that search results would also have been much improved if the participants had had the option to search based on content descriptions or images instead of just words and phrases in the transcripts.	This will be mitigated in the final version of the prototype, which will optimally integrate a final version of face recognition (now executed on a more representative set of content) and content description algorithms from WP2.

Table 7: Second evaluation round impact on the development of the final MeMAD integrated prototype.

10 Future prototype development and evaluation plan and dissemination activities

Looking ahead to the development of the final version of the MeMAD prototype, we aim first of all to bring the final version of currently available functionality in place. Thanks to this second evaluation round, we have gathered to-the-point feedback on which parts can be improved, how this can be done, and where the biggest pain points have been identified. These changes will include:

- Improvements to the backend services for ASR, NER, face recognition, audio classification and (provisionally) language detection, etc., provided by WP2-5, but also the way they are integrated into the prototype platform (e.g., by incorporating more features for depicting and post-editing data delivered by the ASR and NER backend services);
- Improvements to the user interfaces for subtitling and content retrieval, where the evaluation has proven that some easy fixes can already improve the usability of the platform, while more intricate filtering mechanisms will need to be put in place to really realize efficient content retrieval;
- Improvements to the integration with expert editing tools to improve video editor usability and metadata exchange capabilities, either using a more optimized way of delivering data to the Avid editing environment, or by integrating with other representative editing software such as Adobe Premiere;
- For interlingual subtitling, further development of an MT system optimized for subtitling is being carried out. In WP6 and WP4 development will continue of multimodal (T4.1) and discourse-aware (T4.2) machine translation models as well as speech-to-text translation, using also new datasets provided by YLE to optimize the models for subtitle translation. In particular, work will be carried out on improving the spotting and timing of machine-translated subtitles by incorporating feature available from the Limecraft Flow platform. The effects of this on-going work on the improved MT system for subtitling will be evaluated in further productivity tests of automated subtitle translation and post-editing, as laid out in the next subsection.

Additionally, of course, the final prototype will incorporate work on those user stories and functional epics that have not already been addressed, including; content summarization and auto-generation of editorial content (Epic 6.6), auto-generation and post-editing of content descriptions (Epic 6.10) and the functionality to support consumer adoption of the metadata generated by the MeMAD final platform (Epics 6.1, 6.8 and 6.9). In particular, this covers the following:

- The extensive integration of deep captioning for content description and the adoption of the corresponding end user authoring prototype delivered from T5.4 in WP5;
- The findings and recommendations from T4.4 in WP4 to improve the search interface of the platform for supporting cross-lingual content retrieval;
- The inclusion of the work from T3.2 and T3.3 in WP3 for content summarization and automated story-building.

We will describe the implementation of the final prototype in D6.8.

10.1 Final project year evaluations

Based on observations made in the previous section, and developments scheduled in the final project year of MeMAD, we have drafted the following schedule of evaluations for 2020, laid out in Table 8. This planning contains more evaluations than initially described in the DoA, as we feel additional attention is required for evaluations to fully test new functionality becoming available in the course of the final project year, but also to close the gaps found with the second round of evaluations described in this deliverable (and as mentioned in the previous section). The evaluations for 2020 are also planned to be spread out more throughout the year instead of clustering them at the end of the project, which will allow for more time to process the evaluation results. The evaluations have been scheduled to match milestones of when software components will be made available after delivery, to the extent that these had been determined at the time of the plenary MeMAD consortium meeting in Paris in February 2020. The final deliverable of this work package, D6.9 will report the findings of this extensive third round of evaluations.

Date	Evaluation	Responsible partners
February/March	Epics 6.5 (Editing) and 6.2 (Searching) by members of the MeMAD ECG.	Limecraft
March	Intralingual subtitling with improved Finnish diarization.	YLE
March	(Optional) One-to-one interlingual subtitle translation entirely within the Limecraft Flow platform.	University of Helsinki, YLE
April	Epic 6.5 (Editing) with improved metadata and YLE journalists and members of the MeMAD ECG.	YLE, Limecraft
May	Epic 6.2 (Searching) with improved visual metadata (incl. extensive Face Recognition) and using extensive cross-lingual content retrieval (cf. the work executed in T4.4).	University of Helsinki, YLE and Limecraft
June	Epic 6.11 (subtitling): interlingual subtitling with a comparison of subtitling workflow executed in Flow vs. and end-to-end black-box system developed by the partners of WP4.	University of Helsinki, YLE
June	Epic 6.11 (subtitling): end user (consumer/viewer) evaluation of the MeMAD subtitling results.	University of Helsinki, YLE
September	Epics 6.2 (searching) and Epic 6.10 (auto-generation and correction of content descriptions): evaluation of content retrieval and post-editing of content descriptions.	University of Surrey, YLE
September	Epic 6.10 (in particular, user story 4.2.1 – content consumption with autogenerated audio and content descriptions) with end-user consumers.	University of Surrey, YLE

September	Epic s 6.1 and 6.8 (searching for and consuming semantically enriched content).	EURECOM, YLE
September, October	Epic 6.6 (auto-generation of stories from archived or ingested content), tentative.	EURECOM, YLE, Limecraft

Table 8: Provisional final project year prototype evaluation agenda.

Finally, we note that in addition to the evaluation calendar presented here, there will be a cross-work package collaboration with WP7 to evaluate the system using a set of real-life Proofs-of-Concept of the MeMAD technology, integrated with stakeholder's infrastructure, including that of consortium partners (YLE) and ECG member organizations. The conclusions from these trials will be reported in D7.4: "Proof of concept and feedback report".

10.2 Dissemination activities

The following dissemination activities have taken place for the second MeMAD prototype and the evaluations described in this deliverable.

- **Presentation:** 31/01/2019: EBU PTS 2019: EBU Production Technology Seminar, Geneva, Switzerland: "A.I.-assisted automation for end user media production – lessons learnt" by Dieter Van Rijsselbergen.
- **Presentation:** 08/02/2019: University of Antwerp and Medianet Vlaanderen: Open Forum 2019, Antwerp, Belgium: "A.I.-assisted automation of Subtitling and Localisation – Lessons Learned" by Maarten Verwaest.
- **Demonstration:** 08-11/04/2019: NAB 2019, Las Vegas, NV, USA: *Demonstration of the subtitling and translation developed partially for the MeMAD project*, which won a best product of the year award at NAB 2019 (as a precursor of potential of MeMAD applications. Cf. <https://www.limecraft.com/2019/04/11/limecrafts-subtitler-named-product-of-the-year-at-nab-2019/>) by Maarten Verwaest.
- **Demonstration:** 11-13/06/2019: EBU MDN 2019: EBU Metadata Developer Network Workshop, Geneva, Switzerland: *Demonstration of the MeMAD prototype* by Simon Debacq.
- **Presentation:** 24-25/06/2019: LT-Innovate Industry Summit 2019 "Where Language Intelligence Meets Business", Brussels, Belgium: "Maximising Exposure of Audio-visual Content Through Automatic Localisation" by Maarten Verwaest.
- **Demonstration:** 13-17/09/2019: IBC 2019, Amsterdam, The Netherlands, *Demonstration of the MeMAD project prototype*, by Dieter Van Rijsselbergen and Maarten Verwaest.
- **Presentation and demonstration:** 18-19/10/2019: European Language Grid Metaforum 2019, Brussels, Belgium: "The MeMAD project in a Nutshell" and demonstration of the MeMAD prototype between presentation sessions, by Dieter Van Rijsselbergen.

11 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] M. van Someren, Y. Barnard and J. Sandberg, *The think aloud method: A practical guide to modelling cognitive processes*, Academic Press, 1994.
- [5] 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.
- [6] K. A. Ericsson and H. A. Simon, *Protocol analysis: Verbal reports as data*, The MIT Press, 1993.
- [7] 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.
- [8] J. Rubin and D. Chisnell, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, Indianapolis, IN: Wiley, 2nd edn., 2008.

Appendix A Epic 6.5 – Editing assistance using multi-modal and multi-lingual metadata:

Evaluation tasks and participant editing script briefing, think-aloud instructions and post-evaluation interview.

Evaluation script for UC2.1 video editing

set A (migration - unemployment - climate change)

Tarkoituksena on tehdä noin 2 minuutin kooste materiaalista, mukaan seuraavat aiheet:

- maahanmuutto pitäisi nähdä Euroopan yhteisenä haasteena (useita puhujia)
 - Vestager: yhteinen turvapaikkaratkaisu
- solidaarisuus
 - Timmermans:
 - solidaarisuutta tarvitaan ihmisten pelastamiseksi hukkumiselta Välimerellä
 - aivan kuten rakennerahasto auttoi Tšekkiä ja oli merkki solidaarisuudesta, meidän tulee auttaa eteläistä Eurooppaa
 - Keller: puolustamme eurooppalaista solidaarisuutta
- tuet ja miten niitä pitäisi muuttaa (ympäristötekniologia, maatalous)

Lisäksi:

- Lisää Nico Cuén ranskankielinen kommentti maahanmuutosta: "Euroopan vastuullisuus on tärkeää" tai "solidaarisuuden Eurooppa"

IN ENGLISH

- migration should be seen as a common European challenge (many speakers)
 - Vestager: Common asylum solution
- solidarity
 - Timmermans:
 - Solidarity is needed to save people from drowning in the Mediterranean Sea.
 - Just as structural funds on Czech Republic were helpful and a sign of solidarity, we should help southern Europe.
 - Ska Keller: We stand for European solidarity
- subsidising things, what should be changed (environment tech, farming)

For Finnish editors:

Add French-speaking quote by Nico Cué on immigration: "European responsibility is important" or on "Europe of solidarity"

For non-Finnish editors:

Add from the Finnish studio discussion after the debate a quote, where Jan Zahradil's rhetoric of repeating about EU minding its own businesses and nation states doing their own decisions is discussed.

set B (tax havens - external politics - euroskepticism)

Tarkoituksena on tehdä noin 2 minuutin kooste materiaalista, mukaan seuraavat aiheet:

- läpinäkyvyys koskee kaikkia
 - Ska Keller: myös suuryrityksiä
- minimivero kaikille yrityksille, vain kansalliset verot vai verokilpailu poikkeuksin
 - Jan Zahradil: valtioilla on oikeus verottaa yrityksiä, EU ei ole valtio
 - Weber: verokilpailu kyllä, mutta digitaalista verotusta varten tarvitsemme yhteisen ratkaisun
- arvopohjainen kauppa: käydään kauppaa sellaisten maiden kanssa, jotka jakavat arvomme, esimerkin näyttäminen, ilmastokysymysten ja ihmisoikeuksien pitäisi olla osa kauppasopimuksia
 - Vestager: ei ainoastaan vapaakauppaa, vaan arvoihin perustuvaa kauppaa
- enemmistöpäätökset
 - Manfred Weber:
 - Eurooppa ei pysty toimimaan, meidän pitää olla yhtenäisiä, siksi enemmistöpäätöstä tarvitaan
 - Eurooppa on liian hidas ottamaan kantaa Libyan ja Venezuelan tapahtumiin

Lisäksi: Lisää Nico Cuén ranskankielinen kommentti kauppasopimuksista:

“vapaakauppasopimukset heikentävät Eurooppaa, koska emme voi pakottamaan kumppaneita seuraamaan standardejamme”

IN ENGLISH

- transparency goes for everyone
 - Ska Keller: this goes for also big companies
- minimum tax for all businesses, national taxes only or national tax competition with exceptions
 - Jan Zahradil: states have the right to tax businesses, EU is not a state
 - Weber: tax competition yes, but for digital taxation we need something shared
- value based trading: trade with countries that share our values, show example, climate and human rights should be part of trade agreements
 - Vestager: not only free trade, also trade in a value based way
- majority vote on decisions
 - Manfred Weber:
 - Europe cannot act at the moment, We need to be united and that is why majority vote would be needed
 - Europe too slow to take a stance on Libya, Venezuela

For Finnish editors:

Add French-speaking quote by Nico Cué on trade deals: “free trade treaties weaken europe because we can’t impose standards on our partners”

For non-Finnish editors:

Add from the Finnish studio discussion after the debate a quote, where Manfred Weber's talks about economy and climate change are discussed.

Think aloud protocol for UC2.1 - editing

The purpose of a think-aloud protocol is to elicit data from the participant regarding their processing of a task: what they were thinking, what potential problems they encountered, how they solved the problems etc. In a TAP, the participant is instructed to verbalise their thoughts outloud as they carry out a task given to them.

Instructions for the experimenter to keep in mind:

- Instructions must be the same (verbatim) for all participants. See script (Eng + Fin) below. If you give the participant any other instructions in addition to that, they need to be recorded somehow.
- Thinking aloud while doing something else is usually an unfamiliar task for most people, so it is essential to give the participants a chance to practice and "warm up" before actual data collection. It is recommendable to combine this with a couple of very short, very easy tasks similar to what they will be doing in the real experiment.
- If the participant is silent for an extended time during the task, the experimenter can prompt them with "keep talking" or similar, but otherwise **DO NOT ENGAGE IN CONVERSATION, DO NOT GIVE FURTHER INSTRUCTIONS** etc.
- Verbalisations should preferably be done in the participant's mother tongue (or language they are comfortable speaking); when dealing with multilingual data, the participant may end up producing partly multilingual verbalisations - this is not a problem as such, but may make the analysis part more complex.
- Remember to record the whole think-aloud from start to finish! Start the recording before you show the participant the actual task and make sure it runs until the participant indicates they are finished.

Instructions to participant

(based on think-aloud instructions in Ericsson & Simon 1993, p. 378):

In this experiment we are also interested in what you think about when you do the editing tasks I am going to ask you to do. In order to do this I am going to ask you to think aloud as you work on the editing task. What I mean by think aloud is that I want you to tell me everything you are thinking from the time you first see script until the video is finished. I would like you to talk aloud constantly from the time I give you the script until the video is ready. I don't want you to try to plan what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time, I will ask you to talk. Do you understand what I want you to do?

Suomeksi:

Tässä kokeessa haluamme tietää, mitä ajattelet kun teet niitä editointitehtäviä, jotka annan sinulle. Jotta saamme siitä tietoa, pyydän sinua ajattelemaan ääneen samalla, kun teet tehtävää. Ääneen ajattelemisella tarkoitan, että haluan sinun kertovan ihan kaiken, mitä ajattelet, alkaen siitä kun näet videon käsikirjoituksen siihen asti kun video on valmis. Tarkoitus on, että puhut ääneen ihan koko ajan siitä kun saat varsinaisen tehtävän siihen asti että video on valmis. Älä yritä suunnitella mitä sanot tai selittää minulle, mitä tarkoitat. Toimi niin kuin olisit yksin huoneessa ja puhuisit itsellesi. Tärkeintä on, että jatkat puhumista. Jos olet pitemmän aikaa hiljaa, muistutan sinua puhumaan. Ymmärrätkö mitä on tarkoitus tehdä?

Interview script (English version below)

Mikä yleisfiilis editointitehtävistä?

Miksi?

Oliko mitään positiivista/negatiivista? (Yleisfiiliksen perusteella)

Mitkä piirteet erityisesti vaikuttivat editointiin (hyvässä ja huonossa)?

Huomasitko, että datan tai litteraation laadussa olisi ollut eroja?

Millainen editointiprosessisi on yleensä tällaisen aineiston kanssa?

Miten puheentunnistuksen/konekäännöksen käyttö vaikutti omaan työprosessiin?

Entä muu metadata?

Voisitko kuvitella käyttäväsi konekäännöstä/puheentunnistusta työvälineenä?

Mitä/miten pitäisi kehittää/parantaa?

How did the editing tasks feel?

Why is that?

Was there anything positive/negative about the tasks? (Based on first answer; if their feelings are negative, ask about anything positive.)

What features of the provided data and the content impacted the editing the most, in good and bad?

Did you notice any differences in the data or transcripts?

What is your editing process usually like with content like this?

How did the use of MT/ASR impact your own work process?

And the other metadata?

Could you imagine using MT/ASR as a tool?

How should it be improved?

Appendix B Epic 6.2 – Searching and browsing for ingested and archived content:

**Evaluation tasks and participant briefing,
post-evaluation interview and participant background
information form.**

Participant briefing, UC2.2

Go through this briefing at the start of the test situation, before starting on the tasks.

1. Make sure the participant has filled out the [background information form](#); if they haven't, have them do it before you begin. Give them an identifier they will use for the feedback forms as well.
2. Explain the purpose of the research: to study the usefulness and usability of automatically generated metadata in searching audiovisual materials.
3. Explain how the data collection situation will proceed:
During this session, you will search for specific things from a limited archive of audiovisual material. You will have two different sets of metadata available for the search. You will search for the same thing first with legacy metadata only and then with legacy metadata + ASR + NER + MT. After each task is completed, you will be asked to fill out a form about the task, with a focus on the latter search (with automatically generated metadata). At the end of the session we will carry out a short interview, which is recorded.
4. If you feel like you aren't getting anywhere with a search despite your best effort, move on to the next search (or task).
5. Time will be limited to a maximum of 20 minutes per task (one task is 2 searches). We will do as many tasks as there is time for. There are six tasks in total.
6. VERBATIM: "In this experiment we are also interested in what you think about when you do the search tasks I am going to ask you to do. In order to do this I am going to ask you to think aloud as you work on finding the suitable clips. What I mean by think aloud is that I want you to tell me everything you are thinking from the time you first see the question until you have found the clip you think is suitable. I would like you to talk aloud constantly from the time I give you the question until the you are ready. I don't want you to try to plan what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time, I will ask you to talk. Do you understand what I want you to do?"
7. Show the participant the platform and how to switch from one set of metadata to another, remind them not to remove pre-existing filters.

Start recording audio when everything is ready; give the participant their first task.

Tasks:

1. **A program where François Hollande gives a speech (official speech at his office room)** [Basic search for a specific program]
2. **Programs that are local / national election debates from the 2014 European Parliament elections. 2 programs from Finland and 2 from France.** [Basic search on program type]
3. **3 programs / clips where wind power is discussed** [Programs / clips that discuss topic X [must be specific enough]].
4. **Where British politician Daniel Hannan talks about immigration** [Clips with person A talking about topic X]
5. **A debate / discussion shot in an outside setting in Moscow** [Clips from an event X / time period X]
6. **Program / clip with horses appearing** [Clips with X appearing [object / place / action etc.]]
7.
 - a. [a map of Belgium as a graphic]
 - b. [2 short clips of people staring at screens (mobile, tablets, tv, computers etc.)]
 - c. [Clips with Italian politicians?]
 - d. [A clip drone footage from Paris]

Naming the tasks for the feedback form:

[participant identifier][task number]

Think aloud protocol for UC2.2

The purpose of a think-aloud protocol is to elicit data from the participant regarding their processing of a task: what they were thinking, what potential problems they encountered, how they solved the problems etc. In a TAP, the participant is instructed to verbalise their thoughts outloud as they carry out a task given to them.

Instructions for the experimenter to keep in mind:

- Instructions must be the same (verbatim) for all participants. See script (Eng + Fin) below. If you give the participant any other instructions in addition to that, they need to be recorded somehow.
- Thinking aloud while doing something else is usually an unfamiliar task for most people, so it is essential to give the participants a chance to practice and "warm up" before actual data collection. It is recommendable to combine this with a couple of very short, very easy tasks similar to what they will be doing in the real experiment.
- If the participant is silent for an extended time during the task, the experimenter can prompt them with "keep talking" or similar, but otherwise **DO NOT ENGAGE IN CONVERSATION, DO NOT GIVE FURTHER INSTRUCTIONS** etc.
- Verbalisations should preferably be done in the participant's mother tongue (or language they are comfortable speaking); when dealing with multilingual data, the participant may end up producing partly multilingual verbalisations - this is not a problem as such, but may make the analysis part more complex.
- Remember to record the whole think-aloud from start to finish! Start the recording before you show the participant the actual task and make sure it runs until the participant indicates they are finished.

Instructions to participant

(based on think-aloud instructions in Ericsson & Simon 1993, p. 378):

In this experiment we are also interested in what you think about when you do the search tasks I am going to ask you to do. In order to do this I am going to ask you to think aloud as you work on finding the suitable clips. What I mean by think aloud is that I want you to tell me everything you are thinking from the time you first see the question until you have found the clip you think is suitable. I would like you to talk aloud constantly from the time I give you the question until the you are ready. I don't want you to try to plan what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time, I will ask you to talk. Do you understand what I want you to do?

Suomeksi:

Tässä kokeessa haluamme tietää, mitä ajattelet kun teet niitä hakutehtäviä, jotka annan sinulle. Jotta saamme siitä tietoa, pyydän sinua ajattelemaan ääneen samalla, kun teet tehtävää. Ääneen ajattelemisella tarkoitan, että haluan sinun kertovan ihan kaiken, mitä ajattelet, alkaen siitä kun näet kysymyksen siihen asti kun olet löytänyt sopivan klipin. Tarkoitus on, että puhut ääneen ihan koko ajan siitä kun saat varsinaisen tehtävän siihen asti että video on valmis. Älä yritä suunnitella mitä sanot tai selittää minulle, mitä tarkoitat. Toimi niin kuin olisit yksin huoneessa ja puhuisit itsellesi. Tärkeintä on, että jatkat puhumista. Jos olet pitemmän aikaa hiljaa, muistutan sinua puhumaan. Ymmärrätkö mitä on tarkoitus tehdä?

Interview script

Interview to be recorded after all tasks are completed. Finnish version below.

What is your overall feeling about the search tasks?

(Why is that?)

Was there anything positive/negative about the tasks? (Based on first answer; if their feelings are negative, ask about anything positive.)

What features in the metadata impacted the search tasks the most?

Did you notice any differences in the metadata quality?

How did the use of automatically generated metadata impact your own search process?

Could you imagine using this kind of metadata in your work?

How should it be improved?

Mikä yleisfiilis hakutehtävistä?

(Miksi?)

Oliko mitään positiivista/negatiivista? (Yleisfiiliksen perusteella)

Mitkä piirteet metadatatassa erityisesti vaikuttivat hakuun?

Huomasitko, että automaattisen metadatan laadussa olisi ollut eroja?

Miten automaattinen metadata vaikutti omaan hakuprosessiisi?

Voisitko kuvitella käyttäväsi tällaista metadataa työssäsi?

Miten pitäisi kehittää/parantaa?

Background information form UC2.2

Data protection/privacy statement

This form is used to collect background information from participants in the evaluation of tools developed in the MeMAD project. Information will not be distributed to parties outside of the MeMAD 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 Kaisa Vitikainen (kaisa.vitikainen@yle.fi).

* Required

1. Identifier *

Enter the participant identifier provided by the tester

2. Research information sheet *

Mark only one oval.

☐ I have received sufficient information about the research and handling of my personal information within the project. I have received a research information sheet for research participants and data protection/privacy statement.

3. Consent *

Mark only one oval.

☐ I have understood the information provided to me and wish to participate in the research.

4. I may be contacted for follow-up studies *

Mark only one oval.

☐ Yes

☐ No

5. Age *

Mark only one oval.

☐ Under 25

☐ 25 - 39

☐ 40 - 54

☐ 55 or over

6. Gender *

Mark only one oval.

- ☐ Female
- ☐ Male
- ☐ Other
- ☐ I prefer not to answer

7. Language skills *

Which of the following languages do you know?

Check all that apply.

- ☐ Finnish
- ☐ Swedish
- ☐ French
- ☐ English

8. Content description familiarity *

How familiar are you with content descriptions used at Yle and/or INA?

Mark only one oval.

	1	2	3	4	5	
Not at all familiar	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very familiar

9. How much experience do you have in searching audiovisual content (eg. tv series or movies)? *

Mark only one oval.

	1	2	3	4	5	
No experience	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very much experience

10. Have you used automatically generated metadata for searching before? *

Mark only one oval.

- ☐ Yes, often
- ☐ Yes, sometimes
- ☐ Yes, once or twice
- ☐ Never

11. Please answer the following statements *

Mark only one oval per row.

	Completely disagree	Somewhat disagree	No opinion / neither agree nor disagree	Somewhat agree	Completely agree
I keep up with technical developments in my field	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like to test new technological tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New tools make my work easier	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning new tools takes too much time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

This content is neither created nor endorsed by Google.

Google Forms

**Appendix C Epic 6.11 – Intra- and interlingual subtitling:
Participant briefing, post-evaluation interview and
participant background information form.**

Participant briefing

(English version next page)

1. Muistutetaan osallistujaa, että tarkoituksena ei ole pyrkiä korvaamaan kääntäjää, vaan selvittää, olisiko konekäännöksestä apuvälineeksi kääntäjälle.
2. Kerrotaan, miten tilanne etenee:
Tässä koetilanteessa on tarkoitus tehdä kaikkiaan kuusi lyhyttä, noin 3 minuutin pituista klippiä. Kaksi klippiä tehdään alusta asti itse, neljässä on pohjana konekäännös, jota editoidaan. Tarkoitus on tehdä aina kokonainen klippi yhdellä istumalla. Yksittäisen klipin työstämiselle ei ole varsinaista aikarajaa vaan työskentele omaan tahtiin (joskin käytännössä kokonaisaika on rajallinen). Jälkieditoitavien klippien jälkeen täytetään lyhyt kyselylomake. Välissä voidaan pitää pientä hengähdystaukoa. Ihan lopuksi vielä lyhyt haastattelu, joka äänitetään.
3. Muista, että konekäännöksissä aina virheitä, niistä ei kannata hermostua. Samat virheet saattavat toistua. Virheistä näkee, että ne eivät ole ihmiselle tyypillisiä.
4. Jälkieditoititehtävissä korjaa vain sen verran kuin on tarvis, älä jää viilaamaan. Etsitään “riittävää” tasoa, ei täydellisyyttä. Jos et ole varma, onko pakko muuttaa, älä muuta.
5. Kokonaan itse tehtävissä hae vastaavaa “riittävää” tasoa, älä jää viilaamaan. Hoidetaan tehtävä pois alta. Jos jokin kohta on erityisen haastava, esim. puheesta ei tahdo saada selvää, älä kuitenkaan jää kovin pitkäksi aikaa sitä pohtimaan. Jos et ole tyytyväinen ratkaisuusi, laita repliikkiin dollarimerkki \$.
6. Voit käyttää nettiä yms. vapaasti, kuten tekisit normityön ohessa
7. Älä sulje mitään ohjelmista missään välissä.
8. Kun saat klipin valmiiksi, tee uusi repliikki ja kirjoita siihen VALMIS, ja kerro valvojalle.

1. Explain the purpose of the research; remind the participant that the goal is not to replace translators, but rather investigate whether MT could be a tool for the translator (in AV translation context).
2. Explain how the data collection situation will proceed:
During this session, you will work on a total of six short clips, approximately 3 minutes each. Two of the clips will be subtitled “from scratch” without MT, and for four clips you will have a raw version created with MT which you will post-edit. The idea is to subtitle each clip in one sitting. There is no strict time limit for working on each clip, you can work at your own pace (although in practice, the time for the whole session is limited). After post-editing a clip, you will also fill in a short questionnaire about the experience. We can take short breaks in between. At the end of the session, we will also carry out a short interview, which is recorded.
3. Explain to the participant that MT may contain errors, and the same errors may be repeated, try not to get irritated by those. Some of the errors may also be of the type that a human would not make.
4. When post-editing the MT, correct only things that have to be corrected, try not to spend too much time on polishing the translation. The goal is to find a “good enough” level, not “perfection”. Rule of thumb: If you are not sure if something needs to be changed, don’t change it.
5. When doing the subtitling from scratch, aim for a similar “good enough” level without too much polishing. The goal is to just finish the translation. If you encounter a particular problem, for example, it is difficult to make out what the speakers are saying, don’t get stuck for too on that spot. If you feel you cannot find a satisfactory solution, indicate such subtitle with a dollar sign \$.
6. You can use the internet and other resources as you normally would when subtitling.
7. Don’t close any of the programs at any point during the test.
8. When you finish a clip, add a new subtitle and write READY, then inform the tester.

Interview script (English version below)

Mikä yleisfiilis jälkieditoititehtävistä?

Miksi?

Oliko mitään positiivista/negatiivista? (Yleisfiiliksen perusteella)

Mitkä piirteet erityisesti vaikuttivat editointiin (hyvässä ja huonossa)?

Huomasitko, että konekäännöksissä olisi ollut eroja?

Millainen käännösprosessisi on yleensä tällaisten lyhyiden klippien kanssa?

Miten puheentunnistuksen/konekäännöksen käyttö pohjana vaikutti omaan työprosessiin?

Voisitko kuvitella käyttäväsi konekäännöstä/puheentunnistusta työvälineenä?

Mitä/miten pitäisi kehittää/parantaa?

How did the post editing tasks feel?

Why is that?

Was there anything positive/negative about the tasks? (Based on first answer; if their feelings are negative, ask about anything positive.)

What features of the MT/ASR impacted the editing the most, in good and bad?

Did you notice any differences in the MTs?

What is your subtitling process usually like with short clips like these?

How did the use of MT/ASR impact your own work process?

Could you imagine using MT/ASR as a tool?

How should it be improved?

Background information form UC4

Data protection/privacy statement

This form is used to collect background information from participants in the evaluation of tools developed in the MeMAD project. Information will not be distributed to parties outside of the MeMAD 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 Kaisa Vitikainen (kaisa.vitikainen@yle.fi).

* Required

1. Identifier *

Enter the participant identifier provided by the tester

2. Research information sheet *

Mark only one oval.

☐ I have received sufficient information about the research and handling of my personal information within the project. I have received a research information sheet for research participants and data protection/privacy statement.

3. Consent *

Mark only one oval.

☐ I have understood the information provided to me and wish to participate in the research.

4. I may be contacted for follow-up studies *

Mark only one oval.

☐ Yes

☐ No

5. Age *

Mark only one oval.

☐ Under 25

☐ 25 - 39

☐ 40 - 54

☐ 55 or over

6. Gender *

Mark only one oval.

- ☐ Female
- ☐ Male
- ☐ Other
- ☐ I prefer not to answer

7. How many years have you worked as a subtitler? *

8. Which subtitling software do you currently use most? *

Mark only one oval.

- ☐ Spot
- ☐ Swift
- ☐ TextYle
- ☐ Tempo
- ☐ Q4
- ☐ Eztitles
- ☐ Other: _____

9. Which other subtitling software have you used? *

Check all that apply.

- ☐ Spot
- ☐ Swift
- ☐ TextYle
- ☐ Tempo
- ☐ Q4
- ☐ Eztitles
- ☐ ScanTitling

Other: ☐ _____

10. Which features do you appreciate in a subtitling software? *

11. Have you used machine translation as a tool for audiovisual translation? *

Mark only one oval.

- ☐ Yes, often
- ☐ Yes, sometimes
- ☐ Yes, once or twice
- ☐ Never

12. Have you used machine translation for other purposes? *

For example Google or Bing online translators, automatic translations on Facebook etc.

Mark only one oval.

- ☐ Yes, often
- ☐ Yes, sometimes
- ☐ Yes, once or twice
- ☐ Never

13. Have you used automatic speech recognition as a tool for subtitling? *

Mark only one oval.

- ☐ Yes, often
- ☐ Yes, sometimes
- ☐ Yes, once or twice
- ☐ Never

14. Have you used automatic speech recognition for other purposes? *

For example dictating a text message, voice commands for Alexa/Siri/Cortana etc.

Mark only one oval.

- ☐ Yes, often
- ☐ Yes, sometimes
- ☐ Yes, once or twice
- ☐ Never

15. Please answer the following statements *

Mark only one oval per row.

	Completely disagree	Somewhat disagree	No opinion / neither agree nor disagree	Somewhat agree	Completely agree
I keep up with technical developments in my field	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like to test new technological tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New tools make my work easier	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning new tools takes too much time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

This content is neither created nor endorsed by Google.

Google Forms

Appendix D Participant introduction guide:
Searching and browsing in the MeMAD prototype
(Epic 6.2, User Stories 2.2.*)

Contents

1	Searching and Exploring in the MeMAD prototype (Flow).....	3
1.1	The library and general search interface	3
1.2	Complex searches.....	7
1.3	Exploring search results in-depth	8
1.4	Evaluation-specific setup for Epic 6.3 / User Stories 2.2.* evaluations	10
1.4.1	Accessing the prototype platform	10
1.4.2	Available metadata	10
1.4.3	Setting the metadata selection.....	11

1 Searching and Exploring in the MeMAD prototype (Flow)

1.1 The library and general search interface

Within the Limecraft Flow GUI, media content overviews are available in the library view, onto which users land when they select and open a production. Filtering and searching will result in different results being shown in the library section of the GUI.

The simplest way to search for items is to enter search terms in the search text box, e.g, when entering “hollande” the result view is updated to show only items that have ‘hollande’ appearing somewhere in their metadata.

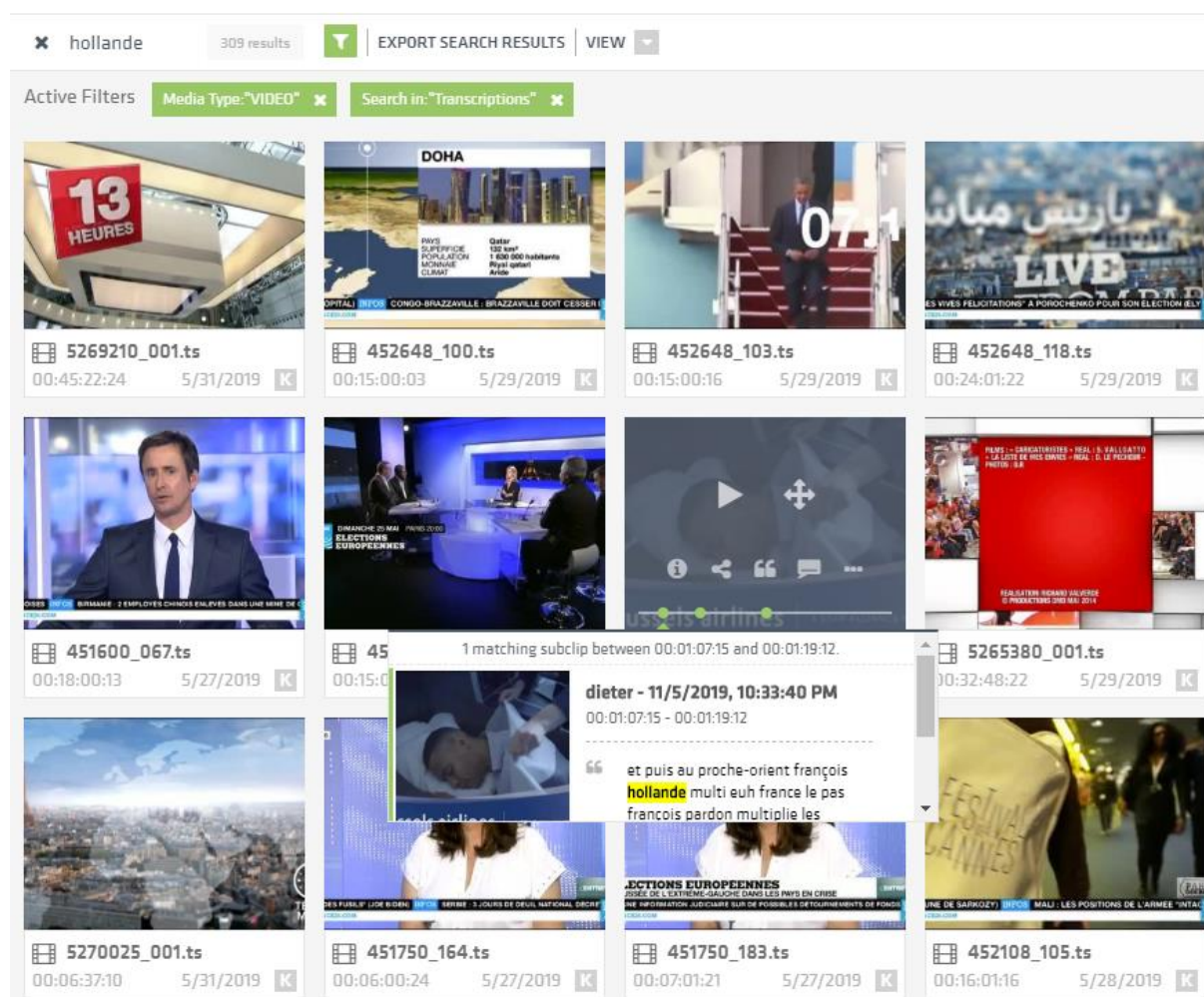
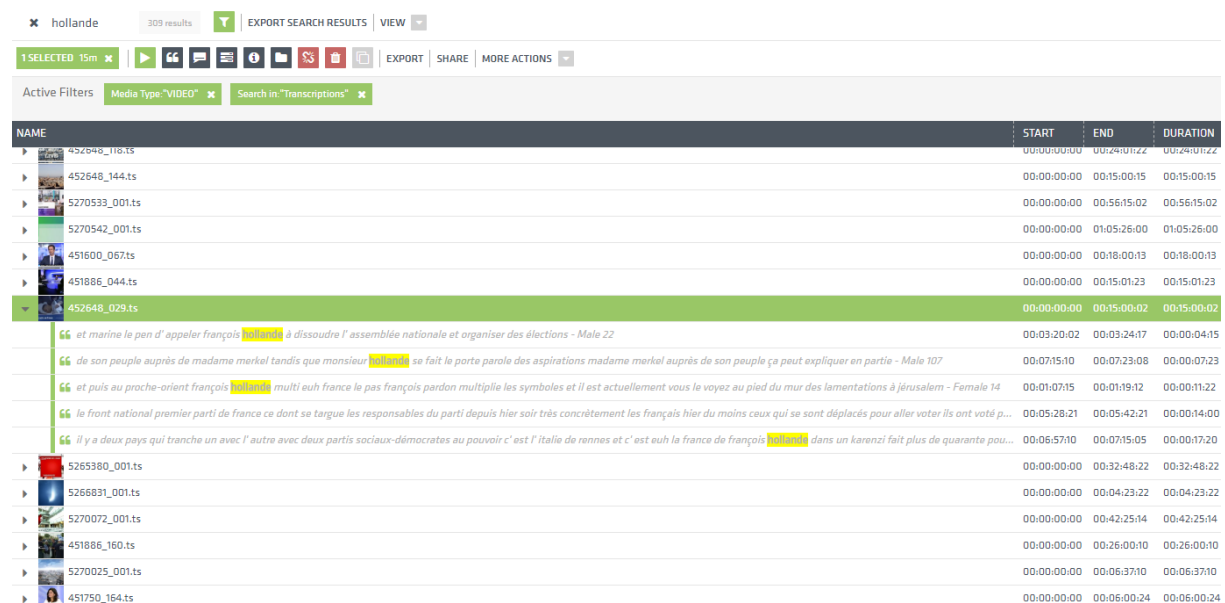


Figure 1: The platform library shows search results for the term "hollande", with clip and audio transcript part matches in the search results.

In the thumbnail view of the library, as shown in Figure 1, a single icon for each clip that matches the search criteria is shown. In case that the query was found in temporal metadata such as audio transcripts and logging information, the timing information of each of the matching metadata elements are used to indicate where (along the clip's

timeline) the metadata fragment was found in the clip. Additionally, highlighting information is also returned by the search index such that the correct element can be visualized in the hovering element (i.e., the text “hollande” is highlighted in yellow as part of an audio transcript).

Alternatively, search results can also be visualized using a list view (cf. Figure 2). Specifically, when results need to be compared across clips, this can help in obtaining a better overview of matching clips, especially as it also allows the display of additional metadata columns, which provide even more insights which ‘matching’ clips are actually relevant for users. Additionally, temporal metadata matches are displayed in the table, along with highlighted matches. Toggling between both views can be done through the “View” drop-down menu (Grid vs. List). From that same menu, users can also select the columns they wish to display for each list row.



NAME	START	END	DURATION
452648_118.ts	00:00:00:00	00:24:01:22	00:24:01:22
452648_144.ts	00:00:00:00	00:15:00:15	00:15:00:15
5270533_001.ts	00:00:00:00	00:56:15:02	00:56:15:02
5270542_001.ts	00:00:00:00	01:05:26:00	01:05:26:00
451600_067.ts	00:00:00:00	00:18:00:13	00:18:00:13
451886_044.ts	00:00:00:00	00:15:01:23	00:15:01:23
452648_029.ts	00:00:00:00	00:15:00:02	00:15:00:02
et marine le pen d' appeler français hollande à dissoudre l' assemblée nationale et organiser des élections - Male 22	00:03:20:02	00:03:24:17	00:00:04:15
de son peuple auprès de madame merkel tandis que monsieur hollande se fait le porte parole des aspirations madame merkel auprès de son peuple ça peut expliquer en partie - Male 107	00:07:15:10	00:07:23:08	00:00:07:23
et puis au proche-orient français hollande multi euh france le pas français pardon multiplie les symboles et il est actuellement vous le voyez au pied du mur des lamentations à jérusalem - Female 14	00:01:07:15	00:01:19:12	00:00:11:22
le front national premier parti de france ce dont se targue les responsables du parti depuis hier soir très concrètement les français hier du moins ceux qui se sont déplacés pour aller voter ils ont voté p...	00:05:28:21	00:05:42:21	00:00:14:00
il y a deux pays qui tranche un avec l' autre avec deux partis sociaux-démocrates au pouvoir c' est l' italie de rennes et c' est euh la france de français hollande dans un karenzi fait plus de quarante pou...	00:06:57:10	00:07:15:05	00:00:17:20
5265380_001.ts	00:00:00:00	00:32:48:22	00:32:48:22
5266831_001.ts	00:00:00:00	00:04:23:22	00:04:23:22
5270072_001.ts	00:00:00:00	00:42:25:14	00:42:25:14
451886_160.ts	00:00:00:00	00:26:00:10	00:26:00:10
5270025_001.ts	00:00:00:00	00:06:37:10	00:06:37:10
451750_164.ts	00:00:00:00	00:06:00:24	00:06:00:24

Figure 2: Search results can also be displayed in a list view, with a breakdown of subclip search results.

To the left of the library’s clip view, user can organize a set of collections and quick views. Collections can be used to store references to clips or subclips, for various organizational purposes, e.g., to indicate the source of material, to catalogue them for logging, for distribution, for story-building, etc. Entire clips or subclips can be assigned to a collection. In any case, the assignment to a collection involves the creation of a reference to the clip; deleting a clip from a collection only removes it from the collection while the clip or subclip itself is not deleted. Clicking a collection (which is then highlighted in green) limits searches to that collection (this selection is also confirmed above the search bar).

Quick Views can be used to store search and display preferences for later use. When the current view (including the selected collection, viewing preferences and activated search queried) is stored as a Quick View, users can later reset this view by activating the Quick

View. We use this mechanism to guide users through the evaluation sessions in Flow (cf. subsection 1.4.3).

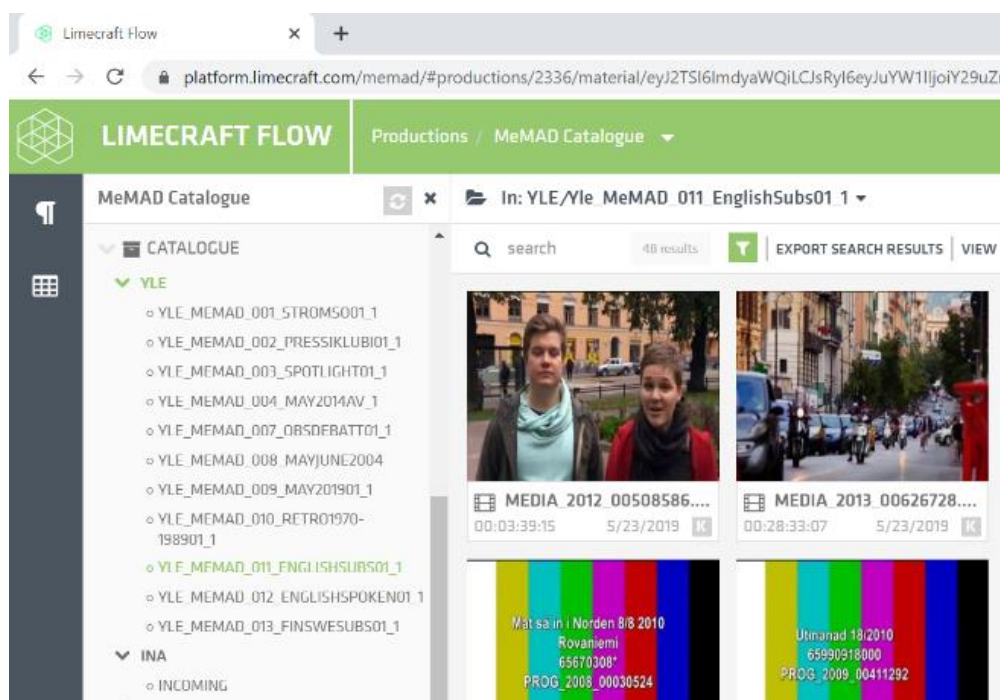


Figure 3: Collections are a means of organizing clips and subclips.

To help users further drill down on results after initial search results are returned based on a textual query, the platform’s GUI provides hints and user interface elements designed to apply further relevant filters.

The first way the search interface helps users is by providing suggestions based on input already typed in the search box. After a small delay, users are presented a list of elements that match what was currently typed into the search box, grouped by facets such as custom metadata fields or entity types (as is the case in Figure 4, which shows auto-suggestions for the *Place* category of named entities).

More significantly, users will also receive suggestions based on partially typed labels of disambiguated named entities that were identified during the clip’s enrichment process. These suggestions will be made to complete any of the partially entered language labels that were stored for the give entity. For example, if the entity “Mediterranean Sea” was identified through the Finnish source text “Välimeri”, users can search for this concept in any language for which the label was stored, such as the German “Mittelmeer”. This translated label would also appear as an auto-completed suggestion.

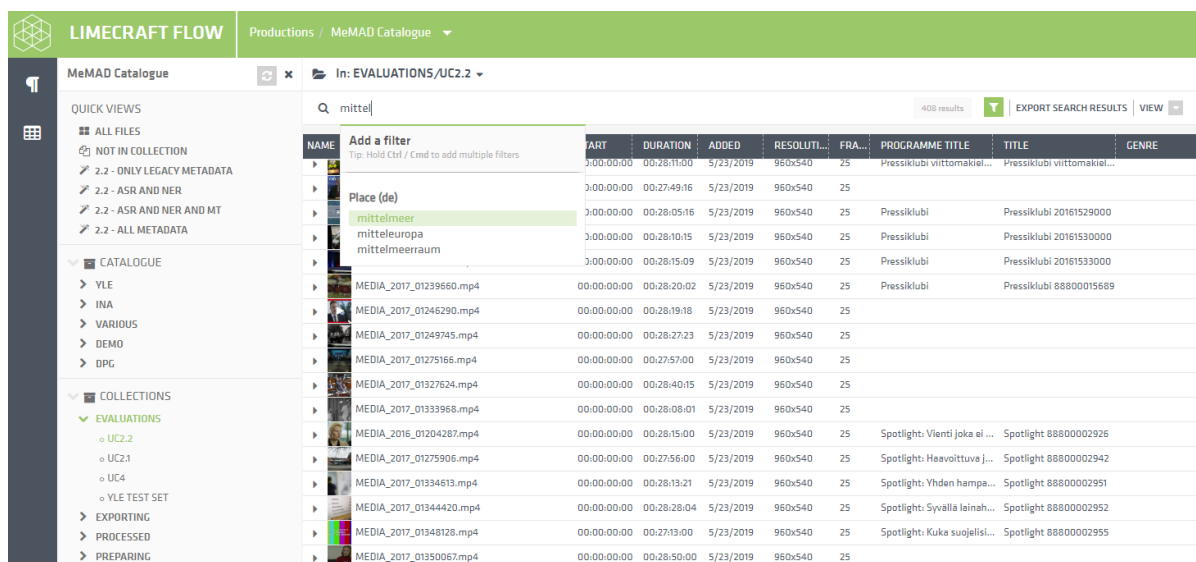


Figure 4: Grouped auto-suggestions using preliminary user input.

Furthermore, users can use the faceted filtering screen to drill down on search results, depicted in Figure 5, including using the following selections:

- Activating which types of metadata to search in, incl. clip metadata, audio transcripts, subtitles.
- Filtering clips that have reached a certain stage in the production process, incl. being logged or having an audio transcript or subtitles.
- Filtering clips using commonly used elements of metadata, such as tags, a rating, types of content (audio, video, documents), or which users created the clip.

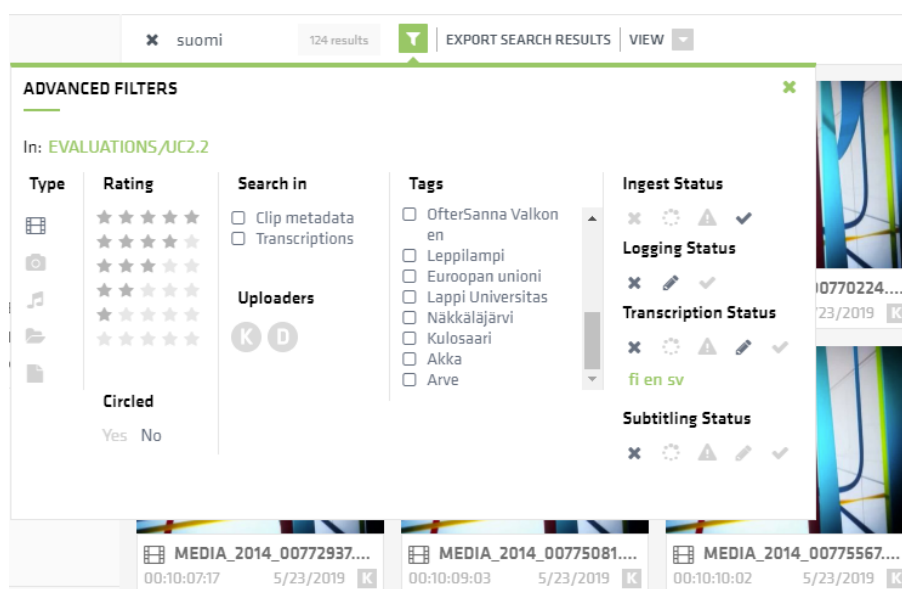


Figure 5: Faceted search panel to further drill down on results using common content properties.

Combining multiple filters and a search query is possible using the platform's user interface. The status of the current active filters are displayed below the search box. The search status can be reset and the filters are removed when clicking the ✕ icon in the search bar. Filters can also be individually removed using the same button on the filter itself (cf. Figure 6).

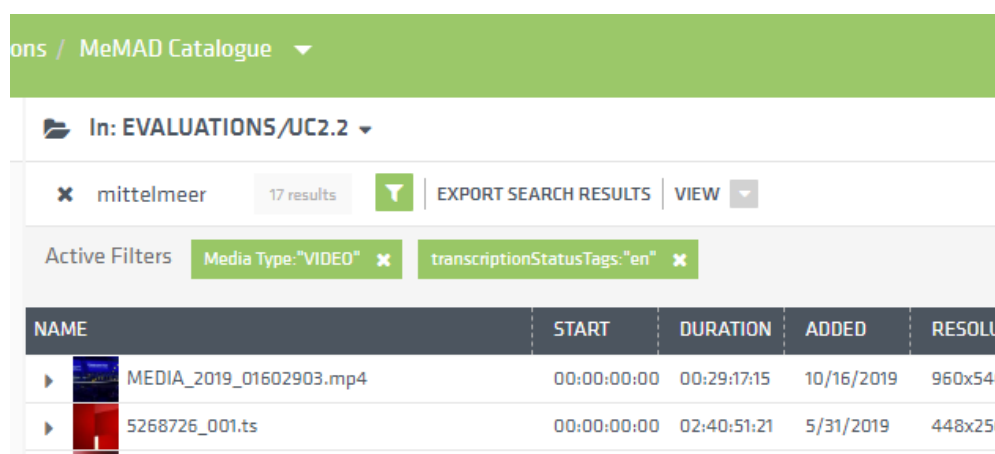


Figure 6: Filters can be toggled and removed when searching.

1.2 Complex searches

The platform supports more complex search queries than literal search strings; criteria can be combined and negated, or filtering can be applied to individual fields of metadata to craft more exact searches. The search functionality has been implemented such that not only generic searches are possible, but also very audio-visually relevant lookup operations can be performed, e.g., with searches on frame rates, clip durations, etc. The following Table 1 illustrates the kinds of searches that can be performed through the search text input field.

Search Query	Description
+good +shot	Search for clips and subclips which have both the word “good” and “shot” somewhere in their metadata.
+”good shot”	Search for clips and subclips which have the literal term “good shot” somewhere in their metadata.
+shot -bad	Search for clips and subclips which have the term “shot” and do NOT have the term “bad” in their metadata.
goo*	Search for clips and subclips which have words starting with “goo” in them.
+tag:nick -tag:pedro	Search for clips and subclips which have the tag “nick” but not the tag “pedro”
+type:AUDIO +tag:briefing	Search for audio clips with the tag “briefing”

name:First*	Search for clips which have a name starting with "First".
framerate:30000/1001	Search for clips having frame rate of 29.97 fps.
camera:ARRI	Search for material shot with ARRI camera
+country:(egypt france)	Search for clips which have a custom field "country" set to either Egypt or France (assuming a custom field with name or label "Country" exists). This functionality is also used for supporting detected named entities. Each category of a named entity is given its custom field, e.g., "Place", "Person", etc., and users can specifically search for entities within their categories.
transcriptionStatusTags:(EDITING;* COMPLETED) +europe	Look up items that contain the term "europe" somewhere in its metadata and that have an audio transcript available (either complete, or being edited at the moment of searching).
+firstPublicationTime:[2015-01 TO 2015-12] +suomi	Lookup up items that were first published in the course of 2015 (i.e., a range between January 2015 and December 2015), and that have "Suomi" in their metadata.

Table 1: Examples of complex search queries supported by the MeMAD prototype platform.

1.3 Exploring search results in-depth

Once users have obtained an initial set of search results, they can explore these results in-depth using other tools provided by the platform. The selection of one or more search results can be done, either:

- Selecting (multiple) individual clips by holding down the Shift or Ctrl keys;
- Selecting all results using Ctrl-A shortcut;
- Selecting clips using drag-and-drop of a selection from within one clip's thumbnail to another.

When a selection has been loaded, specific tools can be loaded with this selection. This includes the transcript tool (shortcut: t), the general clip information tool (shortcut: C) or the subclip logging tool (shortcut: S). This is shown in Figure 7.

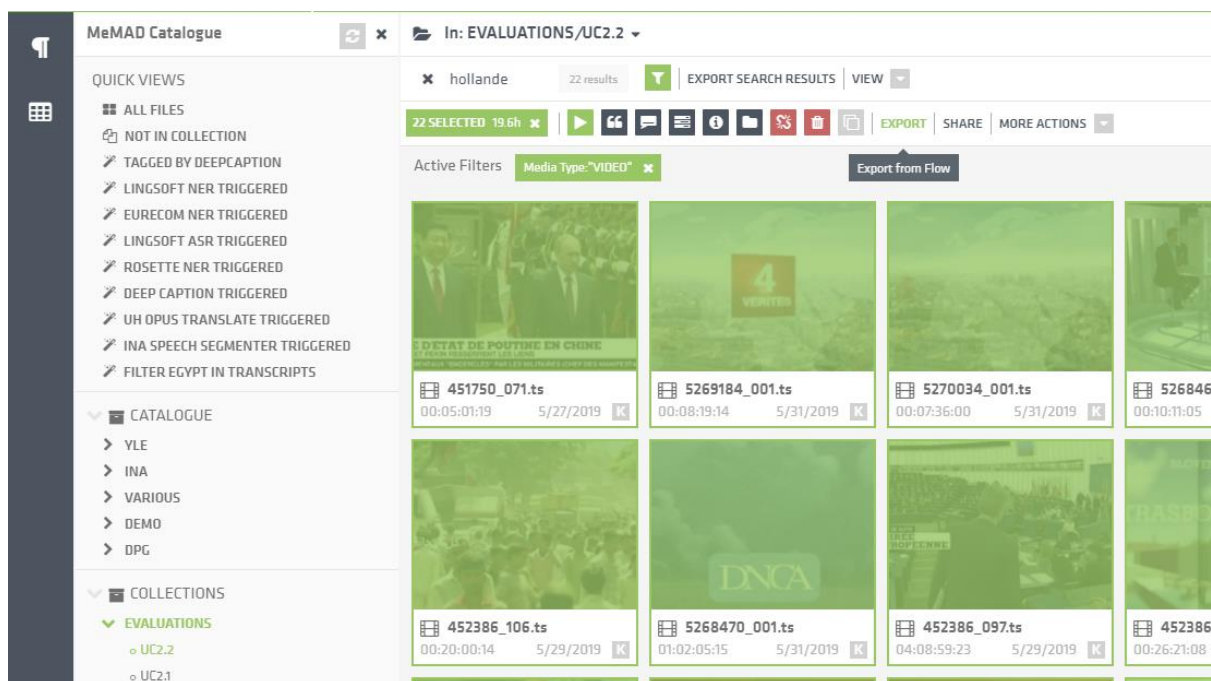


Figure 7: Triggering actions from arbitrary selections of content within search results.

When loaded, the editing tools provide access to the selected clip's details, cf. Figure 8. The different tools can be toggled using the icons on the top right and the previous and next clips from the selection can be navigated to using the left and right arrows (or shortcuts F1 and F7).

Note that the metadata show in the subclip pane can be even further refined using the search functionality if required. Search terms or filters can also be applied here to keep only those subclips that are relevant. In this case, search matches are highlighted in the metadata to further help users find the correct metadata sections.¹

Finally, depending on the data that users are interested in they can toggle different types of subclip metadata on or off at the top of the subclip pane. Users can hide or show e.g., legacy metadata and NER results.

¹ In this version of the prototype, highlighting in subclips is only available for the main field of a subclip. Highlighting does not work yet for translated elements (e.g., the Description (en) part) and highlighting for NER results with multilingual labels only works for the original language labels.

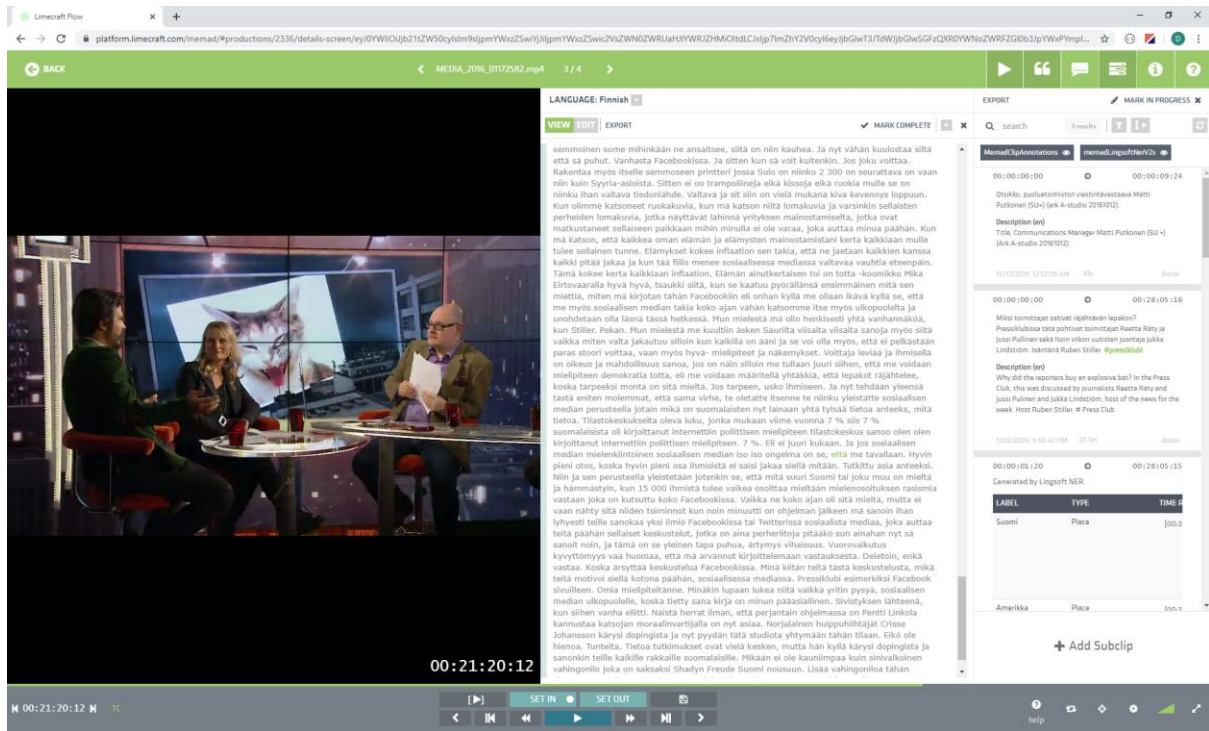


Figure 8: The platform's editing tools with metadata details.

1.4 Evaluation-specific setup for Epic 6.3 / User Stories 2.2.* evaluations

In order to support the formal evaluation of the prototype, we have created a specific setup that can help in configuring the system's user in a predictable way.

1.4.1 Accessing the prototype platform

Accessing the MeMAD prototype platform should be done through the following URL: <https://platform.limecraft.com/memad/>. It is important that this URL is used (with the /memad/ path) to ensure all MeMAD-specific extensions are loaded into the GUI. The production to use is the **MeMAD Catalogue**.

1.4.2 Available metadata

The following metadata is available in the prototype, for all clips in the collection **EVALUATIONS/UC2.2**:

- Most items have their 'legacy metadata' available from the original archive system. This metadata is stored as subclips (under the filter type "MemadClipAnnotation") for the original segments, with an additional subclip to represent the overall description of the clip. Other fields of legacy metadata are available as clip metadata in the info pane. These include:
 - Program and episode title;
 - Genre;
 - Themes;
 - Working title (if applicable);

- Date of first publication/broadcast.

Each piece of legacy metadata has an English translation.

- All but a few items have been audio-transcribed by MeMAD services. This includes items that have the majority of speech in Finnish, Swedish, English and French. All transcripts are directly sourced from the available ASR services and no post-editing has been performed on these transcripts.
- All items with audio transcripts in French and Finnish have been machine-translated into English.
- All items with an audio transcript (in any source language) have been processed by a named entity extraction service.

1.4.3 Setting the metadata selection

To ensure only the correct metadata is searched through for each task, we have created a set of Quick Views for accessing different sets of metadata. Selecting one of the 2.2 - * Quick Views will configure the interface such that only the correct metadata is searched and displayed (Figure 9), and such that the correct media collection is also selected. These quick views are available:

- 2.2 – Only Legacy Metadata: filters down search space to only the legacy metadata of the clip².
- 2.2 – ASR and NER: filters down search space to the audio transcripts and NER results, i.e., the metadata that could be obtained from the audio signal in the audiovisual material.
- 2.2 – ASR and NER and MT: filters down search space to the audio transcripts, NER results and the machine translation of the original language transcript in English. I.e., this is all metadata that was generated automatically, either directly from the source material (first order metadata), or in second order derived from first-order metadata.
- 2.2 – All Metadata: does not filter the search space and allows users to look through all available metadata, as a combination of legacy and auto-generated metadata.

The screenshot shows the LIMECRAFT FLOW interface. The top bar is green with the LIMECRAFT FLOW logo and the text 'Productions / MeMAD Catalogue'. Below this, the 'MeMAD Catalogue' section is active, showing a search bar with '408 results' and a filter button 'Only Legacy Metadata'. The 'QUICK VIEWS' sidebar on the left lists: ALL FILES, NOT IN COLLECTION, 2.2 - ONLY LEGACY METADATA (selected), 2.2 - ASR AND NER, 2.2 - ASR AND NER AND MT, and 2.2 - ALL METADATA. The main content area displays a table of search results with columns NAME, START, and END.

NAME	START	END
MEDIA_2014_00766263.mp4	00:00:00:00	00:27:47:24
MEDIA_2016_01059968.mp4	00:00:00:00	00:27:40:00

Figure 9: Quick Views to support the User Stories 2.2* evaluations.

² Even though the metadata fields *program title*, *episode title*, *genre* and *theme* are legacy metadata, they are available in each of the scenarios.