MEMAD PROJECT: END USER FEEDBACK ON AI IN THE MEDIA PRODUCTION WORKFLOWS

L. Saarikoski[1], D. Van Rijsselbergen[2], M. Hirvonen[3], M. Koponen[4], U. Sulubacak[5], K. Vitikainen[6]

[1,6] Yle, Finland, [2] Limecraft, Belgium, [3] University of Helsinki & University of Tampere, Finland, [4,5] University of Helsinki, Finland

ABSTRACT

This paper discusses the prototypes built and end-user trials run in the European H2020 project MeMAD (Methods for Managing Audiovisual Data) for implementing more efficient media production based on semi-automated media enrichment tools. The prototypes offer automated content annotation supported by machine translation, cross-language search and retrieval of material and automated multi-lingual video subtitling. Alternative evaluation approaches are described for experimental and close-to-production stage use cases, with the focus alternatively on refining the use cases with qualitative methods or measuring productivity with quantitative methods.

Main findings indicate curious user attitudes towards these types of technologies, with current working practices and individual preferences affecting the results quite strongly. Productivity of subtitling and translation work can be improved by incorporating automated speech recognition (ASR), natural language processing (NLP) and machine translation into the workflows. Using large quantities of metadata raises tool UX design questions and is not fully supported by existing tools. For most purposes tested, the users preferred having the additional metadata available, even in lower quality, instead of hiding or discarding low-quality data.

INTRODUCTION

Demonstrations of potential automated metadata extraction services (AME) such as face recognition, automated speech recognition, machine translation and even object detection and scene classification have in the past few years focused on early technical tests or stand-alone user interfaces built to demonstrate the concept. In order to properly evaluate the potential of these deep-learning-based technologies, for which the short-hand term “A.I.” is commonly conveniently used, in media production, the next larger step is to fit these services into existing ecosystems, architectures and workflows in a local context of a media company. This shift from proof-of-concepts (PoC) into production tests marks several important changes, challenges and practical considerations, most notably:

- Envisioned services are for the first time tested in end-to-end workflows instead of isolated sub-processes. Also, the evaluated user experience expands to include all

the parts of the user work process and how different parts of the work tie into each other.

- On top of the technical performance metrics, a layer of more business-oriented success criteria is introduced, such as productivity and user satisfaction.

Typically, also the amount of data increases between iterations in the evolution from proof-of-concept to in-production use, as amounts of content and number of services involved in a workflow increase. Furthermore, at the stage of production tests, the element of optimizing dataflows is present: Out of the large number of AI services, which ones should be combined and what parts of their data output should be used to create an optimal work process?

The European Horizon2020 project MeMAD attempts to research the challenges mentioned above, with research groups developing the algorithms and other core elements of machine learning technologies such as automated speech recognition (ASR), computer vision and machine translation (MT) for audiovisual media data. Building on these, the project pilots the use of these technologies as iterations of a project prototype, and the most promising elements are further evaluated in a close-to-production use by the Finnish Broadcasting Company Yle, the French National Audiovisual Institute INA, and other interested parties.

This paper focuses on the evaluation of the MeMAD technologies with focus on the stakeholder point of view. The project evaluation activities are referred to as a case study, demonstrating the methods and issues that are relevant in the stage of fitting the project technologies into existing professional production workflows. The full project evaluation reports can be found at https://memad.eu and they are summarized in this paper when needed. New evaluation results will be reported throughout the project and this paper describes the findings as of April 2020, shortly after the second of the three project evaluation rounds has finished.

EVALUATION DESIGNS FOR VARYING PURPOSES

Overall aim of these evaluations is to understand the usability of deep learning based technologies in metadata creation and machine translation from the user perspective. The scope of evaluation described here focuses on the user needs and use cases from professional media production.

Most MeMAD technologies are still new to potential users in the creative media production (e.g. editors, journalists, archivists and subtitlers). Therefore, this evaluation 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 new technologies to better assess their practical applicability. The basic approach was to give the participants a hands-on experience, building the test situations so that they resemble authentic production situations. Our approach stems from the usability research and applies the iterative design (see e.g. Tan et al (1)), feeding information from evaluations back to the development of improved MeMAD prototype and technologies.

Thus far the MeMAD evaluation work has been done on three tracks in the media production process:

1. Video editing: How can the video editing process take advantage of AME technologies and the metadata it generates?
2. Media archive searching: How can the content retrieval process, from production databases and archives, be implemented using automatically generated metadata, and to what extent can those metadata replace descriptions input by archivists?
3. Subtitling and translations: How can the subtitling and translation processes be assisted or automated using machine learning technologies?

Like the readiness levels of different technologies, each of these tracks are in different stages of development and 'readiness' in terms of production level workflows and end-user adoption at evaluation participant organizations. This has led to slightly modified evaluation setups for each of the tracks, though there are also shared elements. The different readiness levels provide a good background for discussing alternative evaluation setups. Each of these tracks and their main findings this far are described in more detail below.

The basic setup of all evaluations has been that evaluation participants have performed a task or multiple tasks similar to their everyday work and evaluation data has been gathered during and after the evaluation session.

- For the video editing track, the task was to edit a summary mash-up from a longer program, in this case, an EU-elections debate.
- For the media archive track, the tasks were variations of typical archive searches of varying types.
- For the subtitling and translations track, the tasks included the creation of same-language (interlingual) and translated (interlingual) subtitles.

By varying the data available and workflow design used for each task, alternative approaches were explored and as a result, these alternatives can be compared to each other.

For the data collection, we combined qualitative and more quantitative approaches and gathered data from users performing controlled tasks with the following methods: Think-Aloud Protocols (see: van Someren et al (2)) or process data (keylogging) during the tasks, and User Experience Questionnaire (UEQ) (see: Laugwitz et al (3)) and brief semi-structured interviews after the tasks were completed. A modified version of UEQ was used, with 13 adjective pairs on a 7-point scale (e.g. practical - impractical). The interview transcripts were analyzed thematically (Matthews and Ross (4)).

The qualitative methods produce data on subjective evaluations by the users and what potential problems they encountered and how they solved the problems. In the think-aloud protocol, the participants are asked to verbalize their thoughts out loud and this is recorded and analyzed afterwards. 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 of the task execution and keystroke measurements when testing the subtitling processes.

For the video editing and media archive searching tracks, the qualitative approach was favored as these workflows were still in the stage of refining the use cases and validating their potential. Timing the task performance and measuring keystrokes and mouse activity

were seen as premature for these cases, and more weight was given to the impressions and opinions of participants with the idea of using this input to iteratively find the optimal use cases for MeMAD technologies.

In comparison, the subtitling and translations track focused more on quantitative data - keylogging process data and UEQ - as the use cases for this track are more mature and the evaluation served the purpose of optimizing the work process rather than drafting alternatives for the work process design to be further explored.

The main technical platform used in the evaluation process was the MeMAD prototype based on the Limecraft Flow platform. However, the prototype user interface was used only for the media archive searching track, where the participants' everyday working environments differed from each other and the prototype platform was used as a neutral user interface to avoid bias on how familiar the evaluation system was to each participant. For the other two tracks, the evaluation tasks were performed using the same production tools the participants normally use, incl. video editing and subtitling software, and the MeMAD prototype was used as a background service to analyze and preprocess the content used in the evaluations.

In the following sections we will go through the main findings and the evaluation setup for each evaluation track.

LOOKING FOR THE SWEET-SPOT APPLICATION: CRAFT VIDEO EDITING

The use of automated metadata and machine translation in video editing suites was the most experimental from our evaluation tracks. When starting, the use case was an educated guess or an assumption. Main goals for this track was to gain a better understanding of this use case, and to validate and test the ideas of providing the video editors with additional multilingual data to help the video editing process.

For this evaluation, a ca. 90-minute recording from the 2019 European Elections lead candidate debate was pre-processed in the MeMAD prototype platform to create transcripts in English, and French, facial recognition data on the participants and machine translations from the French transcripts into English. These inputs were then loaded into the Avid Media Composer editing stations (Figure 1) that the participants used in order to edit a short mash-up video based on a rough script provided to them. The participants were a small group of professional video editors working for Yle.

In terms of the tools, this setup was close to the everyday working environment of the participants. However, the video editors criticized the task for being less detailed than what they typically encounter. This highlights the highly specialized professions involved in media production and underlines the importance of recruiting participants from varied backgrounds in this stage where the exact use cases are still unclear. As an example, the participants commented that the extra metadata they were provided with would in many cases benefit their colleagues browsing through raw footage and writing scripts, but the video editors themselves found only limited use for this type of data.

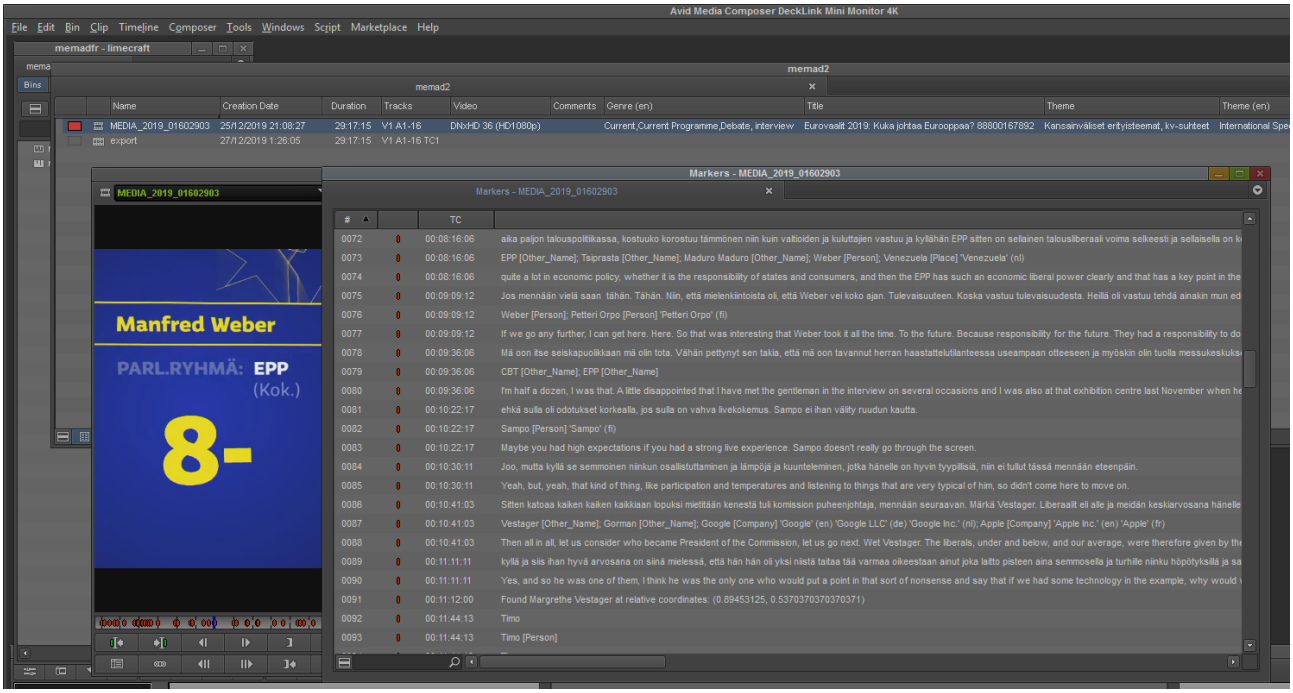

Figure 1 – View of the metadata imported into Avid Media Composer. The metadata is displayed as markers on the clip timeline and in the Markers window.

Another aspect that became apparent was that the existing tools and software, though familiar to the participants, limit the benefits that could be gained from additional data. Since the editing software was not designed to hold and make use of large quantities of time-coded metadata and transcripts, the ways the participants could actually search and make use of the data was limited. For example, even though the participants were provided with time-coded transcripts and facial recognitions, they could not combine these into the same search which would have been needed to perform sub-tasks such as “find the section where person X discusses topic Y”. For follow-up evaluations, we will need to post-process and merge distinct metadata elements into single elements that can be better indexed by editing tools.

This evaluation track demonstrated the idea of piloting new solutions close to production while keeping the setup light, and within the native capabilities of the editing suites used by professionals. This prevented us from investing too much time and effort on e.g. finalizing system integrations or setting up large scale productivity measuring and analysis frameworks before the actual user needs had been validated well enough. The questionnaire and the interview focused on user experience gave the needed insights and ideas to enhance this evaluation track for the next iteration, most notably with a new focus group of more search intensive job descriptions and a metadata set that is structured with the target system’s possible limitations in mind.

VALIDATING THE AMOUNT AND MODALITIES OF DATA: SEARCHING VIDEO ARCHIVES

The media archive track had a clearer use case to start with, as this area had been more thoroughly piloted and investigated in our previous projects (see e.g. Saarikoski and Eaton (5) for details). While the use case of automatically generating and translating metadata to power media findability was clear enough, the evaluation gave valuable insight on which types of media retrieval tasks and which metadata modalities worked well together.

Also, this track included a series of tasks modelled according to real-world examples from the archive services of Yle and INA. Each task involved looking up items in the MeMAD content catalogue of 210 media hours through the Limecraft Flow search interface (Figure 2). The six task types performed were basic searches of a single program, a specific program type, and searches focusing on topics, person and topic combinations, locations and visual objects appearing. Also, the desired outcome of the tasks reflected the production use: instead of one correct answer for each of the tasks, the participants were instructed to aim for a short list of "good enough" candidate clips or programs

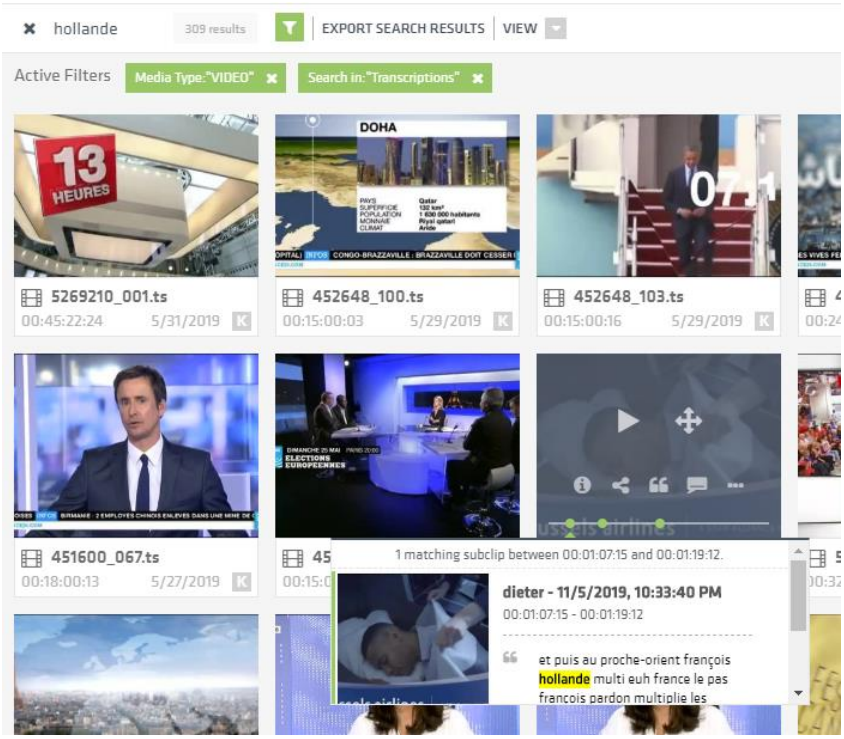

Figure 2 – View of the search interface showing results for the term 'hollande', with clip and audio transcript part matches in the search results.

Each task was performed twice: a) First relying only on the metadata provided as-is from the archives of Yle and INA and b) relying on all metadata, transcripts and machine translations available on the project prototype platform. Each of the content items had been refined with, in addition to the original archive metadata,

- English machine translations of the original metadata (in Finnish and French)
- ASR results in the main source languages (French, Finnish, Swedish, English)
- English machine translations of the ASR results
- NER results based on the ASR output, linked to Wikidata for multilingual labels and descriptions for the NER output

Many of the key findings from this evaluation track this far relate to the amount of metadata that is being handled and displayed. Though the multiple data sources and

modalities provide ample points for searches to hit, they also require functionalities and visual cues to support the users and their navigation within the media collection. Filtering and faceting the metadata and search results, and clearly indicating into which modality each of the results belong to were features mentioned in the user interviews. For example, the same search term could easily be found in the transcripts and visual annotations and this needs to be clear to the users when they browse through the search results.

While all participants expressed interest in using these functionalities in their work and they were also generally positive and interested in future possibilities, there are several things that can still be improved. A similar evaluation will be performed during the last project year, complemented with visual analyses such as facial recognition. After including this missing main modality to the metadata, different combinations of data modalities might be worth looking into in more detail to find out if some of the data provided is redundant or irrelevant and, in fact, additional noise in terms of the searching and browsing.

After some development iterations, this track could most likely be ready for a more productivity-oriented evaluation (see the next evaluation track for details on this). However, some elements in the test setup would need to be changed. Instead of using the project prototype platform, the task should be performed in the participants' normal working environment. This way the participants would be familiar with the functionalities of their search system, instead of still getting familiar with the prototype platform's features. Similar effect applies also to the annotation practices that have been followed in the metadata creation. Either the data and analysis processes should be optimized to resemble the existing metadata the participants typically work with, or the participants should have enough time to explore the metadata they are provided with so that they would know what the database in which they are searching actually contains.

PRODUCTIVITY AND END-USER EXPERIENCE: SUBTITLING AND AV-TRANSLATIONS

The third evaluation track, intra- and interlingual subtitling, serves as an example of a use case which has been refined and validated well enough to focus on actual productivity metrics instead of the more exploratory approach of the other two evaluation tracks described above. The purpose of this evaluation was to determine a) how automatically generated transcripts and subtitles affect the work of intralingual subtitlers and b) how automatic translation of subtitles affects the work of interlingual subtitlers.

As in the other cases, this evaluation setup sought to replicate the normal working environment for the participants both in terms of tools and tasks. The evaluation data was prepared with MeMAD AI technologies and then imported for finishing into the subtitling software the participants normally use. For the intralingual subtitling case, video clips were first treated by ASR to obtain speech transcripts, which were then turned into subtitles using natural language processing (NLP) and configurable spotting rules (e.g., maximum characters per line, minimum subtitle duration, maximum word rate, etc.) to obtain optimal subtitle splitting. In the case of interlingual subtitling, the starting point was manually authored intralingual subtitles, which were machine translated using a translation model trained and optimized on subtitle data corpora.

In this study, professional subtitlers created subtitles for both intralingual and 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.

All participants carried out several tasks, each with 2 media clips ca. 3 minutes long. For each participant, a baseline was established by the task of creating subtitles "from scratch", and the other tasks were variations of ASR or MT tools. The participants subtitled different clips from EU election debates and lifestyle and cultural programs, and the productivity metrics were compared. To account for potential differences related to the difficulty of each clip, the clips and MT outputs were rotated in a round-robin format. Task order was also varied to minimize facilitation effect.

With this type of evaluation setup, we get closer to the metrics we need to decide whether technologies such as ASR and MT are worth the investment. Differences in task times and use of keyboard can be compared, as well as individual differences between the participants. However, the nature of different actions that are seen here as just keystrokes needs to be observed and explained separately. For example, the number of corrections, navigation with arrow keys, punctuation etc. cannot be distinguished from one another based on the metrics, but still they may be meaningful for the participants overall reception of the workflow and technology. Subjective evaluations of the task are best captured through the interviews and questionnaires, resulting e.g. in findings that while editing the ASR output was described as relatively easy, fast and simple, at the same time the experience was characterized as relatively boring and limiting.

To summarize the results of this evaluation track, the process metrics indicate that post-editing ASR or MT can increase productivity in intra- and interlingual subtitling. The use of ASR and MT outputs in subtitling reduced the number of keystrokes needed, and in that way have the potential to increase productivity. Regarding reduced task time, the findings were still inconclusive, which may have been affected by the fact that the participants were not familiar with the task of correcting these outputs (see Figure 3).

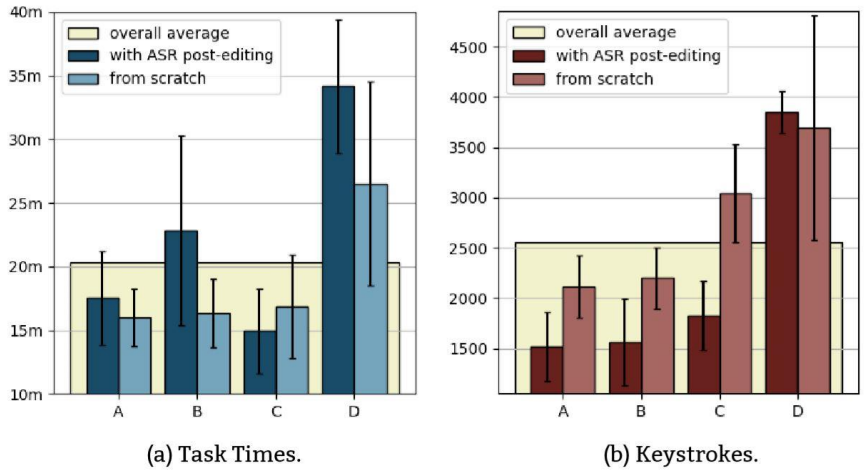

Figure 3 – Average task times (a) and average number of keystrokes (b) when post-editing ASR output (left) and when subtitling from scratch (right) for each participant (labelled as A, B, C, D).

For intralingual subtitling, ASR with post-editing shows promise as a workflow, with most participants indicating they would be interested in using it further. Large differences between participants' task times were observed, and 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 effectively appears a feasible approach, given that we recorded a productivity increase over manual subtitle translations using all variations of MT (Figure 4). This was the case even though the initial MT-based subtitle generation process also inadvertently introduced timing discrepancies that required manual correction by the test panel.

LESSONS LEARNED ON EVALUATION PRACTICALITIES

For a stakeholder thinking whether it makes sense to invest in these technologies, the three evaluation tracks described above provide examples of three different approaches to set up the evaluation, each with a slightly different evaluation focus based on the readiness level of the use case in question.

When focusing on validating and refining the design of the envisioned use cases, the questionnaires, interviews and think-aloud protocol are good tools. However, the amount of analysis needed after the actual evaluation sessions differs greatly, with the questionnaire scores being the lightest to post-process and the think-aloud protocols the heaviest. A rough estimate of the post-process time per participant is 2-3 hours for the interviews and a full working day or more for the think-aloud protocol. For a resource- or time-oriented project this needs to be kept in mind.

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 user group,

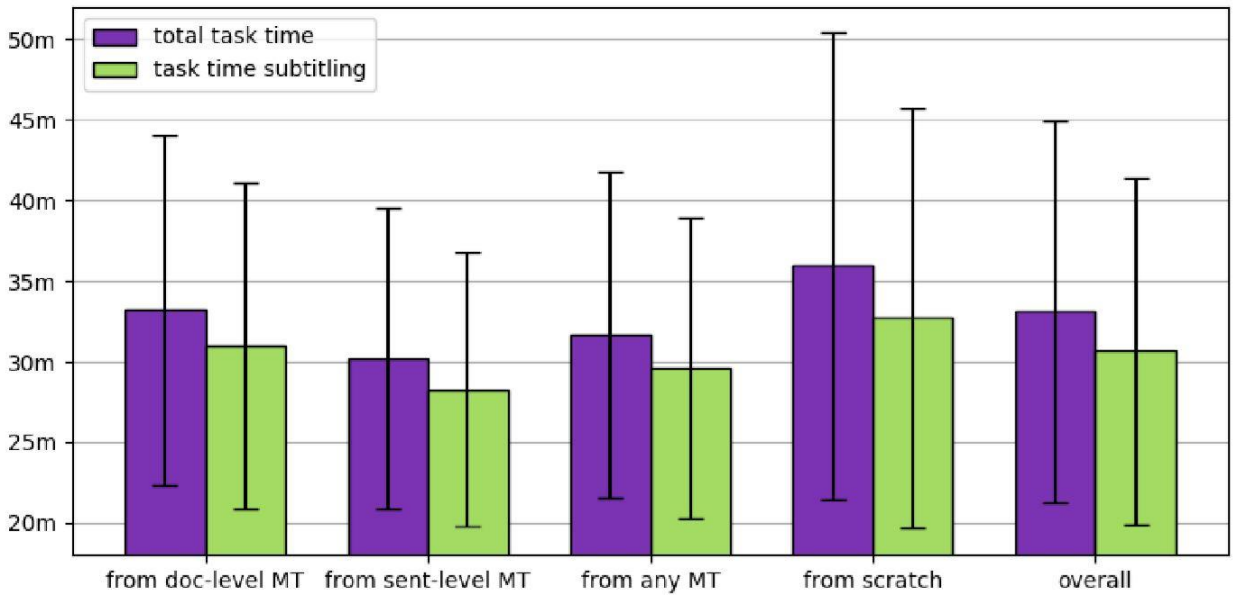

Figure 4 – Average total task times (left) and task times subtitling (right) for interlingual subtitling.

(Rubin and Chisnell (6)). To avoid as much bias as possible stemming from the data or the users, these evaluations should be later expanded to include a wider usership and more content types.

The subtitling track's evaluation was more productivity oriented with a larger number of users. For this purpose, the keylogging, screen capturing, and task time measurement were used with success in this project. However, they also require enough resources to properly post-process the gathered data. The first two evaluation tracks only required tasks demonstrating the new technologies to gather feedback on them. For productivity measurement, an extra task of regular work setup was needed to establish a baseline for the performance metrics.

To control biases in the productivity-oriented evaluation, additional effort is needed to prepare a large enough collection of source media and task variations. Several equally difficult subtitling tasks with equally difficult media clips need to be identified and prepared. The more complicated the task, the more difficult it becomes to control which task and media alternatives actually are comparable with each other. For example, a media search task of finding "person A talking about topic X" would need source media clips with people and topics appearing equally frequently in the evaluation data set, and these people and topics would need to be recognized and annotated with equal quality in the dataset.

Regarding the data needed to start an evaluation, these methods do not require a special ground truth dataset to be prepared (which would be needed if metrics such as word error rate or precision/recall would be calculated). The main approach is to compare different runs of the same task against each other, not to measure the number of "correct" answers against a ground truth dataset.

Which tools and workflows to use is one of the more important variables to consider, if the evaluation goal is to improve existing work setups. Using familiar tools eliminates the need for evaluation participants to learn new tools and their functionalities, and the same goes for the data to be used in the evaluation. A limited test period in production use with real production tasks and data would be optimal as the final iteration of evaluations - for the MeMAD project, this is discussed below.

Another reason to use existing production tools if possible is to expose the current tools and systems to the possible new requirements posed to them by the new types and quantities of data. This way new knowledge can also be gained on which parts of the tools used possibly need to be improved or replaced, as well as what workflow designs could benefit from additional semi-automated metadata.

CONCLUSIONS FROM THE FIRST MEMAD EVALUATIONS AND FUTURE WORK

To summarize the findings from our first MeMAD project evaluations, the overall reception and attitude of the professional communities involved was positive towards the AME and MT solutions tested. Bearing in mind that most of the MeMAD evaluations this far represent early stage prototype applications with a limited number of participants, our findings indicate the following:

1) User reactions and responses in general towards these types of technologies were positive and curious. Current standard production workflows and user roles have a strong effect on user expectations, though. To some extent, a mismatch between

the provided and expected metadata could be observed, especially in the case of archive professionals normally working with conceptually higher-level human assigned metadata, such as keywords, instead of full-text transcripts and low-level tags. Resolving this issue will require both adaptation to such a new content retrieval context on the user's end, and the adoption of more natural language processing technologies for deriving valid concepts from literal metadata.

2) The productivity of subtitling and translation work can be improved by incorporating ASR, NLP and machine translation into the workflows, but this may reflect on the user experience and is dependent on the content type. Variation in individual preferences and productivity effects should still be further investigated. Also, applications and use cases should be chosen appropriately: bringing automated subtitling to subtitling professionals might not be the best approach; the technology should rather be implemented in scenarios where no dedicated subtitlers are available or a (quasi) complete automation of the process is required.

3) User experience and interface design require special attention when dealing with large quantities and multiple modalities of data. Not all professional media tools can handle or present large volumes of data in a sensible way, and for some types of work there is a need to zoom in and out on the data granularity depending on the task. Work remains to find or develop sweet-spot GUIs that are tweaked for optimally including these AME technologies, or to find ways to reduce the amount of content such they can be reasonably visualized by existing specialist tools.

4) If the number of data enrichment tracks that can be produced or presented is limited, using results from ASR and facial recognition seem to be a good starting combination for video editing and archiving tasks, based on our test panel participants' comments.

5) For most purposes tested here, the users preferred the additional metadata available, even in lower quality, instead of hiding the data if the quality is not good enough. Nevertheless, the concern for information 'overload' and 'too much data' were also recorded. As such, AI technology implementers should attempt to find better ways to limit abundance of data whenever possible and present only those data that are relevant to the user in their specific process in the media production chain.

Even though many details still need to be investigated, elements of the prototypes tested could already be incorporated into production systems. In many of the cases, the added value of e.g. automated transcriptions or machine translations was recognized by the participants, and the main criticism targeted the user experience of the systems used or the way data was structured or presented, rather than with the data themselves.

To gain further insight into the potential business value of these applications, a larger, system level evaluation is needed for the searching and subtitling tracks, combining the data creation processes with the user / consumer processes to assess the overall value of these technologies. For the media archive track this means that semi-automated or fully automated metadata creation workflows should be evaluated in connection with the search and browse workflows that make use of the metadata. For the subtitling track the semi-automated subtitling will be evaluated in connection with the end-consumer reactions to

these subtitles to get a more fine-grained understanding of the potential application areas. The remainder of the MeMAD project, in the rest of 2020, will already facilitate an attempt to realize this. The MeMAD evaluation tracks continue and expand their work, and a track focusing on consumer services will be introduced. Evaluations will be expanded into production-oriented PoCs, working on actual live productions to see to what extent the findings from the limited user panels hold up in real-world content creation scenarios. We will also introduce optimizations regarding issues seen from the first evaluations, including improvements to subtitle generation for machine translation, combining more modalities of metadata (incl. topic detection, automated video captioning and audio classification) and devising ways such that legacy professional applications can be better served with the AME-produced metadata.

More information on the on-going evaluations can be found on the project website at https://memad.eu and will also be reported in future conferences and publications.

REFERENCES

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

2. M. van Someren, Y. Barnard and J. Sandberg, 1994. The think aloud method: A practical guide to modelling cognitive processes.

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

4. B. Matthews and L. Ross, 2010. Research Methods: A Practical Guide for the Social Sciences.

5. L. Saarikoski and M. Eaton, 2019. Identifying use cases and evaluating ML technology. Conference presentation at FIAT/IFTA World conference 2019. Presentation available at https://www.slideshare.net/fiatifta/saarikoski-yle-metadata-machine.

6. J. Rubin and D. Chisnell, 2008. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, 2nd edn.

ACKNOWLEDGMENTS

This work is part of the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No 780069).