

EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action Call: H2020-ICT-2018-1 Call topic: ICT-29-2018 A multilingual Next generation Internet Project start: 1 January 2019

Project duration: 36 months

D6.3: User needs and challenges for news media industry (T6.1)

Executive summary

The purpose of this report is to derive and prioritise user needs of the news media industry for the EMBEDDIA project. We analyse innovation challenges in news media and map the current state-ofthe art in media technology with respect to artificial intelligence (AI) tools in newsrooms. To increase our understanding of news media user needs we organised a project workshop where user needs were explored, applying action-research paired with user-oriented design as the methodology. The results from roundtable discussions in the workshop provide guidance for coming development work in the project. The concluding part presents ideas that were most prominently discussed in the workshop. There are the three key recommendations from the workshop about what EMBEDDIA should focus on: comment management, detection of interesting news, and personalized news generation. During the progress of research, this preliminary guidance will continue to be expanded based on what turns out to be important, preferable, and technically feasibile.

Partner in charge: UH

	Project co-funded by the European Commission within Horizon 2020 Dissemination Level	
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-







Deliverable information

	Document administrative information
Project acronym:	EMBEDDIA
Project number:	825153
Deliverable number:	D6.3
Deliverable full title:	User needs and challenges for news media industry
Deliverable short title:	User needs
Document identifier:	EMBEDDIA-D63-UserNeeds-T61-submitted
Lead partner short name:	UH
Report version:	submitted
Report submission date:	30/09/2019
Dissemination level:	PU
Nature:	R = Report
Lead author(s):	Carl-Gustav Lindén (UH-SocS)
Co-author(s):	Marko Milosavljević (UL-SocS)
Status:	_ draft, _ final, <u>X</u> submitted

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
02/04/2019	v1.0	Carl-Gustav Lindén (UH-SocS)	Draft report template
12/04/2019	v1.1	Hannu Toivonen (UH-CS)	Comments to report
05/06/2019	v1.2	Silver Traat (TEXTA)	Comments to report
07/07/2019	v1.3	Marko Milosavljević (UL-SocS)	Addition, state-of-the-art
14/08/2019	v1.4	Carl-Gustav Lindén (UH-SSoc)	Draft report sent for comments
26/08/2019	v1.5	Marko Pranjić (STY)	Comments to report
26/08/2019	v1.6	Salla Salmela (STT)	Comments to report
27/08/2019	v1.7	Carl-Gustav Lindén (UH-SocS)	Implemented the requested changes
28/8/2019	v1.8	Matthew Purver (QMUL)	Additions to report
29/8(2019	v1.9	Senja Pollak (JSI)	Additions to report
30/08/2919	v2.0	Carl-Gustav Lindén (UH-SocS)	Implemented the requested changes.
02/09/2019	v2.1	Carl-Gustav Lindén (UH-SocS)	Report submitted for review
06/09/2019	v2.2	Marko Robnik-Šikonja (UL-CS)	Internal review
09/09/2019	v2.3	Luz Saturnino UEDIN	Internal review
13/09/2019	v2.4	Carl-Gustav Lindén (UH-SocS)	Report submitted for quality check
15/09/2019	v2.5	Nada Lavrač (JSI)	Quality control
17/09/2019	v2.6	Carl-Gustav Lindén (UH-SocS)	Implemented the requested changes.
18/09/2019	final	Nada Lavrač (JSI), Anita Valmarska (JSI)	Final quality check and finalisation
30/09/2019	submitted	Tina Anžič (JSI)	Report submitted

Table of contents

1	Int	roduction	4
	1.1	Main purpose of EMBEDDIA	4
	1.2	Development and adoption of new technology in media and journalism	4
	1.3	Artificial intelligence, news media and journalism	5
2	Ov	verview of new digital tools and assistants in the news media industry	6
3	Us	er needs of EMBEDDIA identified in a workshop	11
	3.1	Workshop methodology: Action-research and user-oriented design	11
	3.	.1.1 Action research	12
	3.	.1.2 User oriented design	13
	3.2	Preparations for the Tallinn workshop	14
	3.3	Methods for gathering data at the workshop	14
	3.4	Results from the workshop	15
4	Co	onclusions	
	4.1	Recommendations	
	4.2	Discussion	
R	efere	ences	24
A	ppen	ndix 1: Participation in Computation + Journalism Symposium	
A	ppen	ndix 2: Questionnaire for survey	

List of abbreviations

AI	Artificial Intelligence
DoA	Description of Action
EC	European Commission
ExM	Ekspress Grupp
NLG	Natural Language Generation
NLP	Natural Language Processing
STT	Finnish News Agency
STY	Styria Media Group
WP	Work Package





1 Introduction

The aim of this report is to derive and prioritise user needs of the news media industry for the EMBEDDIA project. User needs are identified by reviewing the research literature on user needs in news media, exploring already existing technical solutions that are relevant to the project and identifying needs that they address, as well as discussing the findings from our own exploration of user needs of media partners in the project.

The report is divided into four parts. In the introduction, we give a short overview of the EMBEDDIA project and innovation challenges in news media. The second part offers an analysis of the current state-of-the art in media technology with regards to artificial intelligence (AI) tools in newsrooms. User needs were explored in a project workshop which is covered in part three where we present research design, preparations, and action-research with user-oriented design as the preferred methodology. The results from the discussions in the workshop provide the guidance for the coming development work in the project. The report ends with conclusions and recommendations.

1.1 Main purpose of EMBEDDIA

While advanced automated language technology tools and resources exist for a few dominant languages (English, French, German), many of Europe's language communities - and the news media industry that serves them - lack appropriate tools for multilingual internet and text-based industry development. EMBEDDIA aims to solve some of these limitations and develop methods and tools, an EMBEDDIA Media Assistant, to allow to transfer resources and tools from technologically well-supported languages to others. The techniques and technologies developed in the project will enable journalists, editors and researchers to search, link and monitor news reports and editorial content; analyse and react to public user comments; produce content semi-automatically, and to do all this across European languages.The aim is to provide tools for smaller and medium-sized media companies and newsrooms in under-resourced languages that might otherwise lag behind.

Three small and medium-sized European media companies are part of the project. These are Ekspress Grupp from Estonia (ExM), Styria Media Group (STY)¹ from Croatia, and the Finnish News Agency (STT). They have different profiles. ExM is a leading media group in the Baltic states focusing on publishing, printing services, and online media content production. STY is one of the leading media groups in Austria, Croatia, and Slovenia with a broad offering of services ranging from newspapers to books and radio stations. STT is the only news agency in Finland. The publicly available EMBEDDIA Media Asistant will be developed for them and other media companies by the Estonian language technology company Texta.

1.2 Development and adoption of new technology in media and journalism

Investing in new technology and reforming newsrooms around artificial intelligence is costly and strategically complicated (Agrawal et al., 2018). It is less about the costs of investment in software and more of choosing the right editorial strategy and making sure that the newsroom is equipped with the right skills for radical transformation. This becomes a problem specifically for small and medium sized media companies. Due to several reasons, such as the downfall of advertising revenues and changing news consumption habits, the media industry is struggling not just to reinvent the business model but for survival (Lehtisaari et al, 2012). The lack of an innovation culture does not help either. Instead, imitation and copying are core elements of media innovation (Boczkowski, 2005). For instance, a study of US metro papers shows that companies are mostly rearranging existing business models rather than experimenting with radical innovations, while examples of new projects breaking

¹ In EMBEDDIA, STY has been recently replaced by Trikoder, a collaborating company from Croatia, however STY will still be the source of news and provider of other media content and needs for the Croatian language.



the barriers of legacy media are rare (Villi et al., 2019). Especially newspapers tend rather to reproduce the concepts of past successes, focusing on incremental changes, rigorous brand alignment and top down-monitored, commercially steered activities (Järventie-Thesleff, Moisander, and Villi 2014). Media executives often hesitate to make bold, high-risk moves because the landscape keeps changing so fast and the risks with costly investments is high (McDowell 2011). The brakes on change are also cultural (Ess, 2014). Going from products to services, from hardware to software, and from audience to users and consumers includes changing mindsets, many times unlearning the trade and its institutional truths.

1.3 Artificial intelligence, news media and journalism

The emergence of a new technology tends to be viewed as an incremental process, with the new inevitably replacing the old technology. However, examples from history show an interactive process wherein each shapes the other (Chadwick 2013, 25-26). Artificial intelligence provides a technological challenge that is different from previous transformations as it is not just about new hardware, say radio or television, but a whole new human-machine system. Artificial intelligence (AI), automation, and machine learning are transforming journalism around the world. Al is already affecting the news value chain, from news gathering to content processing, creation and distribution (Marconi & Siegman, 2017; Weghe, 2018). New technology is for instance used to verify input sent by users, such as pictures or videos, to filter out unwanted user comments, to check facts, to personalise and enhance the reading experience through content recommendation, to automatically generate texts and video, or to use advance audience metrics based on user behaviour to analyse performance of content. For instance, the New York Times is using predictive analytics tools to gain a competitive edge. This includes so called funnel analysis to see how people become subscribers, and how to influence more to do so. They also use natural language processing to understand what content topics generate reader engagement, helping marketing teams to select articles to promote (Burns, 2015). At BBC an AI tool called Juicer aggregates news and extract content. It watches 850 RSS feeds globally to take and tag articles at scale. Reuters tracer is a large scale system for detecting and verifying real-time news events from Twitter (Liu et al., 2016).

However, journalists struggle to understand what AI means and usually focus on industry announcements. As an example, the most common news source for AI in British news media has been the Silicon Valley entrepreneur Elon Musk (Brennen, Howard, & Nielsen, 2018). Among journalists and media managers, there is frustration about the lack of clarity around key definitions in AI, for instance the connection between intelligence and artificial. The hype around AI does not make the conversation easier: what is the material substance or is this just the new shiny trend of the moment (Perretti, 2019; Milosavljević & Vobič, 2019)? Media companies, constrained by financial and human resources as well as skills, need to think hard about where to invest and innovate. Therefore, the adoption of AI applications needs to be preceded by an editorial strategy with a clear vision and sense of which elements media companies should explore and what they should leave aside (Perretti, 2019). This confusion is certainly not confined solely to journalism. Despite 70 years of discussion, there is still disagreement on how to exactly define artificial intelligence. In this report, we rely on two similar definitions that underline the human side of AI, a "human-centered future" (Diakopoulos, 2019) which come closer to the concept of augmented intelligence first promoted by Doug Engelbart (1962).

- "Al is a scientific discipline, like mathematics or biology. This means that Al is a collection of concepts, problems, and methods for solving them (University of Helsinki/Reaktor, 2018).
- "Most basically, AI is a collection of ideas, technologies, and techniques that relate to a computer system's capacity to perform tasks normally requiring human intelligence" (Brennen et al., 2018).

Research on artificial intelligence in journalism is still a small field with a limited number of articles published. The most recent contribution is a collection of expert opinions edited by Seth Lewis (2019). There are more research on specific features of AI in journalism, for instance texts generated by Natural Language Generation systems (Diakopoulos, 2019; Dörr, 2016; Leppänen, Munezero, Sirén-



Heikel, Granroth-Wilding, & Toivonen, 2017; Linden et al., 2019; Sirén-Heikel, Leppänen, Lindén, & Bäck, 2019), recommendation systems (Beam, 2014; Liang, Lai, & Ku, 2006), bias detection in texts (Ali et al., 2010), content moderation with machine learning (Jiang & Han, 2019; Roberts, 2019), and story-finding and news worthiness or augmented creative writing (Carlson, 2018; Huovelin et al., 2013; Magnusson, Finnäs, & Wallentin, 2016; Plattner, et al., 2017; Zachos et al., 2018) but almost nothing on a crucial aspect, the data that drives AI applications in newsrooms. Skills for working with numbers, large and small data sets, public records, and data visualizations are essential in news organizations today (Boyles & Meyer, 2017; Rogers, Schwabish, & Bowers, 2017; Weber, Engebretsen, & Kennedy, 2018). The literature on newsrooms strategies for data is scarce, for instance when it comes to news text generation, where the unhindered access to structural data is a crucial element of operations (Karlsson, 2019). Some strategies for exploring, evaluating and utilising data have been described in the literature (Linden et al., 2019; Magnusson et al., 2016), for instance with software that enables journalists to encode news events and stories directly as data (Caswell, Russell, & Adair, 2015; Caswell & Dörr, 2019).

Certain specific features of newswork provide a challenge. Looking at the potential for AI in news reporting, Jonathan Stray notes that investigative journalism is a hard problem for AI, due to a number of reasons. One important aspect is the unique character of each investigation and the lack of structured data, which makes it hard to create computer models (Stray, 2019)

2 Overview of new digital tools and assistants in the news media industry

In addition to a background of EMBEDDIA project and the review of the research literature, an overview of the existing AI systems/tools/assistants is needed for an in-depth analysis of news media industry and its needs. In this audit, we review existing work and products that are of interest to the project, mainly relevant technologies and their applications. This part stems from both the newsrooms (representing journalistic and editorial needs) and from the management (usually representing the needs of management, owners and advertising sectors). We note that user needs are shaped by different aspects (e.g., position within the media production circle – news gathering, content production, content promotion, content distribution), different approaches (public sector consisting of Public Service Media, community media and similar; private sector consisting mostly of privately, family owned companies or publicly owned companies, trading at stock exchanges etc.), and the particularities of specific media sectors and industries (print, radio, television, digital, different hybrid forms).

This part of the report explores existing technical solutions that are relevant to the project. We provide an up-to-date overview of the state of the art in the news media industry, offering key information on the existing AI systems/tools/assistants from different larger media companies (from U.S.) such as The Washington Post, Bloomberg, Forbes, and insight into their use of AI tools. It should be underlined that this state-of-the-art overview focuses on Europe and the United States. We are aware of the rapid development of new media technologies in China (Lee, 2018) and will return to the topic as the EMBEDDIA project proceeds.

The overview is based on a combination of research methods:

- Systematic document analysis: gathering, following and analysing of industry news reporting and analysis in specialized and quality media, research centers and projects, media reports and technology reviews.
- Personal interviews: in-depth semi-structured interviews that were conducted with key digital managers and editors at key world media companies that are influencing media industry and

have strong innovation role within the news media, including the Guardian (two interviews in London), the Financial Times (two interviews in London), BBC (two interviews over Skype), Reuters (one interview in London), the New York Times (one interview), and Associated Press (one interview over Skype). The interviews were analysed to provide insight into the managerial, editorial, and journalistic aspects of the AI-related issues in the newsrooms of these media companies, and this forms a focus of a research paper co-written by one of the researchers within this project. The research paper has been published in Journalism (Milosavljević & Vobič, 2019).

 Additional discussion on language-related issues and needs with Language Technology Lead from BBC as one of the largest media companies, well-known for its technological innovative approach and influence on other newsrooms in European Union, Europe and throughout the world, to gain additional insight into specific user needs regarding the language aspects in contemporary newsrooms.

Based on such a triangulation, in Table 1 we gathered relevant information and insight into the state of the art regarding the user needs involving new digital tools and assistants in the news media industry.

Media	System	What it does
Reuters	Lynx Insight	A "cybernetic" newsroom that helps the business publisher producing stories since March 2018 (see Chua, 2018).
	Partnership with Graphiq	Provides news publishers with a wide range of free interactive data visualizations (once embedded on publishers' website, the data visualizations are updated in real time).
Associated Press	Wordsmith from Automated Insights	Writes stories on topics including company earnings and minor league sports (see Marconi & Siegman, 2017).
		Increased the number of articles on earnings reports from 300 per quarter to 3,700.
		Generates minor league and college game stories.
Guardian	Chatbot (on Facebook)	Allows users to pick from US, UK and Australian version of Guardian News, choose a delivery time and get selected news stories everyday via Facebook Messenger.
		Users can pick topics.
		The interface replies to chat messages with related content relevant to the users query.

Table 1.	Overview	of new	digital	tools and	d assistants	in the	news	media	industry
	01011101	of field	aigitai	10013 4110	1 43313141113	in the	110113	meana	maasay



Forbes	Bertie (a new CMS)	Recommends article topics for contributors based on their previous output.			
		Recommends headlines for contributors based on the sentiment of their pieces.			
		Recommends images for contributors.			
		Writes rough versions of articles that contributors can simply polish up.			
		Produces the makings of an article, complete with links to relevant, related articles and images.			
		Uses artificial intelligence to make publishing more efficient for Forbes staff.			
		Uses artificial intelligence to make it as easy as possible for visitors to consume multimedia content.			
		Acts as a thought-starter.			
		Has doubled the number of loyal visitors - people who visit more than once per month - and increased traffic on the site (65 million monthly uniques in November 2018).			
Washington Post	Heliograf	Generates short stories based on structured data about things like election results or Olympics events. Has generated thousands of stories.			
		Promotes articles with a local orientation in topics like political races to readers in specific regions – geotargeting.			
		The Post also uses internal alerts to signal anomalous bits of data. Reporters who see the alert can then determine if there is a bigger story to be written by a human being. During the Olympics they set up alerts on Slack, the workplace messaging system, to inform editors if a result was 10 percent above or below an Olympic world record.			
Wall Street Journal & Dow Jones	AI technology	Transcribes interviews.			
		Helps journalists identify "deep fakes".			



New York Times	AI technology	Personalised newsletters.			
		Helps with comment moderation.			
		Identifies images as it digitizes its archive.			
	Editor	Recognizes semantic tags and learns the most salient parts of an article.			
		Searches through data in real time and extracts information based on requested categories.			
		Makes information more accessible.			
		Simplifies the research process.			
		Provides fast and accurate fact checking.			
	Perspective API	Organizes reader's comments interactively so that viewers can quickly see which ones they may find "toxic".			
Bloomberg	BHIVE (human-	Understands what motivates news consumers.			
	and prototyping	Learns how the audience responds to something new.			
	3.2.4	Tries to understand the "why" and hammers away with more intrinsic motivations that might get lost in data sets.			
		Develops a set of news user profiles.			
		Gives specific attributes to different consumers.			
		Classifies users as either a "News Chaser" (someone who wants to be the first to know breaking news), a "News Connector" (who wants to share stories to connect with others), or an "Opinion Seeker" (who wants to get perspectives that allow him to understand what is happening in the world).			
		Collaborates with editorial staff early on in the process.			
		Allows readers to see the most important stories of the moment without having to click away from the home feed.			
	The Bulletin	Serves the needs of "News Chasers".			



Bloomberg News	Cyborg	Assists reporters in producing thousands of articles on company earnings reports each quarter.
		Dissects financial reports.
		Bloomberg also uses internal alerts to signal anomalous bits of data. Reporters who see the alert can then determine if there is a bigger story to be written by a human being.
Patch	AI technology	Assistance to staff reporters and freelancers. Generates 5 to 10 percent of weekly posts.
El Pais	Google Jigsaw	Machine learning to analyse comments and identify problematic items.
BBC News Labs	Juicer (see Liu et al., 2016)	Watches around 850 global news outlets' RSS feeds and aggregates and extracts news articles from the BBC and outside sources.
		Assigns semantic tags to the stories and organises them in one of four categories: organisations, locations, people, and things.
		Helps journalists find the latest stories.
	Data stringer	Helps journalists subscribe to data sources, and will notify when patterns arise or thresholds are broken.
Yahoo! Sports	Wordsmith from Automated Insights	Generates content (articles, reports, emails) with data from specific sports teams.
Quartz	Bot Studio (within creating a set of automated tools for	Users text in with questions about news events, people, or places, and the app replies with content that it believes will be relevant for them.
	journalists/	Interacts with users.
Inject	Proprietary software	Helps journalists be more creative and find story angles faster.
Belga	Proprietary software	The tool (under construction) focuses on detection of published news from three perspectives: political direction, sentiment and factuality.
		Helps journalists and publishers as well as information and communication professionals to monitor, research, evaluate and monetise news media content.
Tamedia	TaDam, NLP/web scraping	Collects data from more than a thousand webpages with different schedules depending on the relevance, sorts the data into a database that can be searched through an interface (see Plattner et al., 2016).



To sum up and prioritise this overview of the state of the art in the news media industry, we can see that most media outlets, which use new digital tools including AI systems/tools/assistants, use them for:

- General purposes: help in producing stories, either the tools write the stories independently or they propose headlines, topics or images.
- Personalisation: tools used for user personalisation of content (articles, newsletters) including geo-targeting.
- Comment moderation: the mentioned tools are used for comment moderation or analysis and data extraction.
- News search or aggregation: this helps contributors to find interesting topics or angles.
- Other tasks: new digital tools include also data visualization, audience engagement, and transcription.

3 User needs of EMBEDDIA identified in a workshop

To be able to go from general ideas of user needs proposed in the project application to actual and prioritised user needs of the involved media companies, the project group organized a workshop in Tallinn on March 12-13, 2019. In that event, media managers, journalists, and developers from media partners worked with researchers. This was an iterative and collaborative process where the researchers found out about media partners' expectations, initial views and attitudes towards the EMBEDDIA project, and defined the input they craved from the media companies. A difficult part of innovation is learning how to identify "important" needs versus "interesting" ones. The aim was to identify challenges that are specific to the operations of different media partners and prepare specifications documentation. It is a common that people most responsible for the design of digital products and services rarely take into account the users' goals, needs, or motivations. This can result in outcomes that lack a coherent user experience (Cooper, Reimann, & Dubberly, 2003).

The three media partners presented specific problems for which they currently do not have adequate solutions:

- ExM in Estonia is struggling with user comments in several languages for which they do not have moderating tools.
- STY is addressing the data needs of an international media company.
- STT have to make technological multilingual advancements that would strengthen their ties with other Nordic news agencies.

It was important to have the opportunity to discuss these concerns in detail, including the preliminary product vision, understanding what is technically feasible, who are the users and potential users, what are the behaviors, attitudes, aptitudes, motivations, environments, tools, and challenges (Cooper et al., 2003). The workshop also provided media managers the opportunity to share not only the technical challenges, but also the insights of what their businesses are trying to accomplish.

3.1 Workshop methodology: Action-research and user-oriented design

The methodology of the workshop was based on action-research theory (Defrijn et al., 2008) combined with user-oriented design (Veryzer et al., 2005). We combined these two methods as they provide an orientation that fosters a deeper understanding of user needs and what delivers value.



3.1.1 Action research

Action research offers a framework for research collaborations between scholars and practitioners with the clear intention to induce change and improvement of practice (Grubenmann, 2016). Its goal is to serve as a tool for "solving problems experienced by people in their professional community" (Appelgren & Nygren, 2014). This research framework has been labelled a "remarkably inclusive methodology" (Cunningham, 2014, 3) and a "collaborative approach" (Stringer, 1996, 15). The main idea is that only an active participant can understand the dynamics of a specific human system, in this case the complexity of media innovations. Without trying to change it, the system remain invisible for the passive observer (Schein, 1987). In line with Dupagne and Chuan (2019), we argue that journalism and news media would be best served by "being proactive toward Al development". In the workshop, we presented a conceptual and methodological framework for interactively and iteratively clarifying and designing software according to the needs of the media partners. Following Defrijn et al. (2008), in the EMBEDDIA project we apply cycles of planning, acting, observing, and reflection to the development of new Al tools for journalists.



Figure 1. Action research cycle (Source: Defrijn et al., 2008).

The main reason behind this iterative approach is that a problem-driven and solutions-oriented research agenda helps us to orientate towards the specific conditions of the media partners involved in EMBEDDIA and the most relevant issues. The aim of the workshop was to function as an operationalisation of the "Identifying, Informing, Analysing" phase of the cycle.

Furthermore, Grubenmann (2016) have defined three principles of action research:

- The intention to change. This solutions-driven approach distinguishes action research from other types of participatory research.
- Participatory and inclusive research. The goal is to maximize the usefulness of research outcomes for a particular community. Outcomes are local and problem specific. However, in the EMBEDDIA project we go further and believe that research outcomes will be generalizable and/or replicable.
- Developmental research process. Outcomes are answers to community-relevant questions created in negotiation with each other. Knowledge is uncertain and ambiguous, answers



tentative and open to modification. This is especially true in the field of media innovations, where new technology is developed and deployed rapidly in a trial and error fashion with uncertain financial outcomes while editorial resources and skills are scarce.

One specific challenge in action research is the terminology used to discuss key concepts and challenges. People from different disciplines or communities of experts have their own vocabularies for framing and presenting their thoughts. Metaphors play a central role in decision making. How they are actually interpreted and used can have a significant impact on decision outcomes (Bosmajian, 1992; Hibbitts, 1994; Napoli, 1999). Insofar as language does not just describe reality but constitutes reality, metaphors do not just describe but shape and create organizational strategy (Inayatullah et al., 2016).

3.1.2 User oriented design

User oriented design is a powerful tool for answering the most important questions that crop up during the definition and design of a digital product. In the case of EMBEDDIA, we discussed a number of important issues and followed a list of questions that allowed for follow up questions:

- Who are the users?
- What are users trying to accomplish?
- How do users think about what they're trying to accomplish?
- What kind of experiences do users find appealing and rewarding?
- How should the product behave?
- What forms should the product take?
- How will users interact with the product?

Table 2 contains a comparison of user-oriented and traditional approaches to designing products.

Table 2. Traditional and new approach to designing products (Veryzer & Borja de Mozota, 2005, 132;Vredenberg, Isensee, & Righi, 2001, 2).

Traditional Approach	User-Oriented Design
Technology driven	User driven
Component focus	Solutions focus
Limited multidisciplinary cooperation	Multidisciplinary team work
Focus on internals architecture	Focus on externals design
No specialization in user experience	Specialization in user experience
Some competitive focus	Focus on competition
Development prior to user validation	Develop only user validated designs
Product defect view of quality	User view of quality
Limited focus on user measurement	Prime focus on user measurement
Focus on current customers	Focus on current and future customers





3.2 Preparations for the Tallinn workshop

To get the expected results from the workshop some preparation work was carried out in alignment with the initial steps of the action research cycle and user-oriented design. The preparations was divided into four parts:

- 1. Planning the methods employed in the workshop;
- 2. Doing the review of state-of-the-art already presented and another round of interviews at the Computation + Journalism Symposium 2019 (see Appendix 1 with list of interviewees);
- 3. Informing workshop participants of the structure and goals of the workshop, and asking them to think about the following questions:
 - How do media managers, developers and journalists interpret what EMBEDDIA is about to achieve and what practical applications do they imagine?
 - What are the best tools different users (editors, journalists, technical staff, etc) can imagine? What functions should the EMBEDDIA Media Assistant have that will help journalists find and analyse news in multiple languages? What problems could it solve?
 - What editorial tasks could it be used for?
 - What would it require for journalists to use it?
 - What sort of user generated content (e.g., commentaries) do the media house wants on their webpages? What sort of support they want to provide to its users?
 - What sort of information from other languages would be useful (for different users)?
 - In terms of automated news generation: where would the data come from, what sort of stories could be created?
 - Which task would different users desire to automate? Which task different users enjoy and would not want to automate?
- 4. Asking the three media partners to fill out an online survey of their expectations before the workshop (see Appendix 2). The questions were directly matching the objectives and task description of the Work Packages WP3, WP4 and WP5. The printed questionnaires were also available during the discussion, and were one of the materials used in the round table discussion.

3.3 Methods for gathering data at the workshop

In the workshop, we undertook context analysis to i) understand how media partners innovate and their needs, ii) to identify stakeholders involved in the innovation process, and iii) to identify opportunities for success of the EMBEDDIA project. The aim of the workshop was to respond to one specific research question "What are the user needs of media partners?" Discussions during the two days took place at four tables and were led by work package leaders. The discussions were recorded by researchers and transcribed by an external service provider, Rev.com. After that, the transcripts were checked by two research assistants. For this report, we focused on the final presentations where discussions were synthesised and summarised. Post-it notes used during the workshop were also categorised and analysed. The purpose of the workshop was to get input from media companies on both newsroom user needs in the different areas covered by the application - from problematic comments and content analysis to generation of news texts in multiple languages. Some user needs were listed in the application for funding, but the list was far from exhaustive and it was expected that others, more urgent needs, might surface in a workshop, which was an iterative exploration of user needs. Therefore, it was important to involve journalists, developers and other relevant media people. With the help of their input, researchers will be better placed to get the priorities right.

Participants focussed on:

• Finding needs and defining challenges: clarify the long term goasl, map the problem, interview experts (present journalists and developers), decide what questions need to be



solved, what user interfaces journalists and developers are used to, what features they think would be useful.

• Idea sketching: Research existing solutions and inspirational examples, identify desired features from elsewhere, sketch out solutions.

The end result is the list of priorities and potential solutions contained in this report that researchers can reference and ensure that EMBEDDIA will be a mutually successful project for researchers and media companies.

The workshop was a chance for journalists and developers to make an impact on the project outcome. It was an important event for the project, but also a chance to sit together and exchange ideas on what digital tools are needed in newsrooms.

The discussion was guided by the responses to the survey mentioned in Section 3.2:

- What would be the most promising applications?
- What would it need to get there? What are the main obstacles and how can they be overcome?
- Is there any data available which could be used for training machine learning models, e.g., manually labelled categories, comments, etc?

The workshop ended with a study visit to the newsroom of ExM and a discussion with the editor-inchief who detailed the challenges of media businesses in the Baltic states.

Finally, media partners and work package leaders were asked to comment on the results of the draft user needs report. This was part of the iterative approach in action-research. As stated, according to Grubenmann (2016), the knowledge is uncertain and ambiguous, and the answers are tentative and open to modification. Therefore, the outcomes need to be created in negotiation with others.

3.4 Results from the workshop

In this section, we present an analysis of the workshop discussions for each work package. The section starts with an overall description of the terms "end user" and "user needs" and how they were used during the discussions. According to the analysis, the terms referred to three different target groups: journalists and editors, media houses, and consumers of journalistic content. Because of the inconsistencies, we have refrained from using these terms in the analysis, even though they were used extensively in the discussions. Instead, we have tried to be more accurate when describing to whom a specific talking point applies. The exception is WP3 where the words 'user' and 'commenter' are used interchangably.

As the EMBEDDIA project relies on mutual understanding and teamwork of academic institutions and media houses from several countries, it is an imperative that it is clear what the terms 'end user' and 'user needs' entails in the project. If not clarified, there is a risk of confusion and turbulence that might endanger both the workflow and the quality of the project. A lack of precision about the users can lead to a lack of clarity about how the product should behave. This discussion will continue as the project proceeds.

WP3: Cross-lingual Technologies for User-Generated Content

The overall objective in this work package is to apply EMBEDDIA's cross-lingual technologies to understand the reactions of multilingual news audiences, thus helping news media companies to serve better their audiences and acting as a basis to assure fairness and integrity of participants in public internet spaces.

Task T3.1: Cross-lingual Context and Opinion Analysis



(Task description: "This task will develop technology for a range of user comment analyses, including topic modelling, conversation structure and context modelling, sentiment, stance and opinion and effect and information spread measurement.")

The media partners expressed an interest in profiling the properties and opinions of their users, to compare and find both similarities and differences between profiles, as well as map out how they react to different stories. However, using labels from human moderators and even labels from some sites that let commenters flag other users' comments for being inappropriate would lead to rather noisy annotations. By instead sourcing labels across several different sites and different languages could allow for a cross-lingual approach, which might lead to less noisy and more accurate annotation from which the project would be able to learn. It was also noticed that the same content being duplicated and ending up in different forms on Facebook and other news sites is a problem, which needs to be resolved later. Another issue that needs being considered is how to get the data behind users' comments that go viral since Facebook no longer allows scraping data from the site unless the data is from your own page (see Bastos & Walker 2018, Walker et al. 2019).

There was a discussion on annotating comments and visualizing them to see what different functions they represent, and what kind of different content is in them. It would function in the same way as tagging, while at the same time visualizing it. The idea is to visualize it in a 'mind map' that would show how many comments are negative, positive, hate speech, off-topic, or constructive in regards to the topic in question. The main idea behind this would be to create a way to help elevate and understand useful or constructive comments and the positives therein, as they are usually few to come by. The aim is that by creating a mind map, the positives would pop up in the visualization and therefore be easier to acknowledge. In short, identifying comments that are useful to journalists, which is usually the same as comments that are unique and different from the majority. It would also help group comments regarding their topic or viewpoint expressed, identify and track peoples' sentiment or stance towards particular topics, and consequently detect whether there is a bias in particular stories, platforms, or countries.

Somewhat related to the above, the media partners were interested in spotting and tracking when people are doing things that are undesirable or unacceptable in public comments; this could include abusive or hate speech, trolling or spamming. For example, off-topic commenting that seems to be a common occurrence, spamming involves one or many users posting the same religious or political rants on several pages and images. Also related to spamming is the concept of trolling which, according to the discussions, is not an individual expressing a particular viewpoint, but an organization pushing an agenda. The ability to detect these behaviours is very important to the media partners, and overlaps significantly with the task T3.2 below. The main specific outputs of this task that were therefore immediately seen as priorities were:

- Identifying topic and sentiment/stance of comments to allow grouping and visualisation;
- Identifying influential commenters and constructive/positive comments;
- Identifying comments that should be blocked and when stories should be closed for commenting (see T3.2 below).

Task T3.2: Cross-lingual Comment Filtering

(Task description: "This task will develop tools for automatic flagging or filtering of user comments, detection of hate speech and political trolling, attempts to elicit extreme reactions and influence others' opinions.")

It was discussed that when filtering comments, it is primarily about trying to analyze properties of individual comments of individual texts. There can be many different reasons as to why someone would want to filter out a comment, whether it is e.g., insulting people, threatening, libeling, slandering, hate speech, or inciting others to violence. As well as in some cases blocking commenters from using language that is not allowed on a particular site or considered inappropriate for a particular article, although these applications differ slightly conceptually. Overall, it would provide a very good



opportunity for using different languages and cross lingual transfer to build good models, even though it is uncertain how accurate the results would be.

The media partners expressed interest and concern over how to tell a user why a comment was stuck in the filter. There was consensus over not telling the person exactly what was wrong with the comment (in the discussion specific words) because the commenter might only exchange the "bad" word for another in response. Instead, it would be better to present a general reason for why the comment was filtered. However, in the back end it would be useful to see the specific reasons (specific words used) when validating a model. One idea raised was that the user would receive a warning when entering an inappropriate text, prompting them to change it before posting. This, however, sparked controversy with some praising the idea, while others were wary of giving people the impression that they are being monitored. It was also pointed out that people who post "bad" things are the ones that tend to get upset about "political correctness" in the media, and who feel as if they are being told what to say and not to say. A Facebook experiment was referenced, where people that were "thumbed down" a lot were likely to get offended and their behavior worsened; which is a situation that needs to be avoided in the EMBEDDIA project. The media partners recognized that there sometimes is a need to close stories and prevent further commenting right at the moment of publishing or sometimes later on. The project could help in this regard by detecting topics that are likely to be controversial or comment sections that are going toxic. According to the media partners, the situation usually occurs when the content is particularly sensitive, for instance involving the death of someone.

In combination with filtering out inappropriate comments, the media partners were very interested in singling out high quality comments, which are comments that promote a healthy discussion on the topic of the article. These positive comments could then be used as content in the article both as a means to improve communication between the media company and individual commentators, and as a way for newsrooms to get engaged in local communities. The media partners would also be interested in tracking the public perception of that particular article promoted with positive comments from users.

The main specific outputs of this task seen as priorities were:

- Detecting comments that should be blocked or referred to human moderators. This could include many categories: abuse, hate speech, spam, self-promotion, use of the wrong language, trolling behaviour etc;
- Automatic labelling with the reason for these blocking/moderation decisions;
- Detection of users who are particularly likely to generate comments needing blocking/moderation;
- Detection of positive/constructive comments (see T3.1 above);
- Detection of commenters' motivations and biases.

Task T3.3: Report generation from multilingual comments

(Task description: "This task will develop and implement methods for generating human-readable reports, in multiple languages.")

This task depends on the previous two and can only be approached further in the future, saw less discussion than T3.1 and T3.2. However, some useful points emerged:

 Increasing engagement: Careful use of reporting could help increase users' engagement with the news, and journalists' engagement with commenters. By using the output of reports of users' comments in media texts, user engagement would be encouraged, and it might also lead to improved civility in the comments section due to increased awareness of the effects of commenting. By allowing journalists to quickly and comprehensibly understand commenters' points of view, reports would encourage this kind of use.



- Explanations: One useful aspect of reporting could be to provide users with an explanation why they or their comments have been banned.
- Source of content: Reports on user commenting behaviour could help give hints for journalists for future interesting content.

WP4: Cross Lingual Content-analysis

The overall objective of WP4 is the development of tools for analysis of news content across languages, to empower news media consumers, researchers and news media professionals.

Task T4.1: Real-time multilingual news linking

(Task description: "The task will develop tools for linking news stories across languages based on their topic and content.")

The media partners expressed strong interest in an alert system that would be able to detect breaking news. This would be especially valuable in small languages as it can take time for news to travel to big international media published in the major languages. In conjunction with detecting breaking news, the system could also be able to detect trending topics and keywords. What constitutes a 'trending topic' is a subject that has to be on social media (Facebook) to be really trending. Therefore, task 4.1: "Real-time multilingual news linking" cannot be disconnected from social media analysis. The system could further be developed to understand how local news spreads into important news. One of the most interesting tools identified by ExM would be a tool to detect important news from neighbouring countries.

There are two different news linking types the media partners were interested in: recommended news and linked news:

- Recommended news refers to connecting other news, with a similar topic to what a person is already reading. STY and ExM have automated recommendation systems, which can serve as a baseline. Since STY is already planning to create similarity news triplet data sets (the process has started since the workshop), it could also be used in improving the recommender (together with tags and keywords).
- Linked news stories refers to articles that function as a background material for the main article (stories that are usually strongly related to the article – the same person or the same discussion covered). These connections are usually done manually by journalists (the case of STY and ExM), but could possibly be turned into an automatic task. Since the workshop, EMBEDDIA partners have already access to the collection of recommended news.

There was concern over how to eliminate overlap in the case a (multilingual) news linking/detection tool is developed. The media partners do not want to burden their readers with an overload of information. The elimination of too much story overlap is vital. Another concern is how to create a robust system that automatically only link to sources that are trustworthy.

Another discussion covered the tags, which are assigned to the articles. STY has already significantly reduced their tagset, while in ExM has tags as well, but no cleaning of the tagset was performed.

Task 4.2: Cross-lingual news summarisation and visualisation

(Task description: "This task will develop textual and visual language-independent multi-document news summarisation.")

Media partners did not find news summarisation and visualisation as one of their priorities. However, when connected to Task 4.1 "Real-time multilingual news linking" this became more interesting. For instance, it would be a useful function in a situation where a journalist finds an interesting topic with connected news in a different language. The journalist could then get a summary of what connected news are focusing on for a better overview of the topic.



However, when discussing summarization it was pointed out that there are three types of summarization that need to be differentiated:

- Summarizing a particular news piece.
- Summarising a group of news, e.g., several multilingual news pieces into one language.
- Summarizing comments, features, reportage etc. (teaser effect).

In conclusion: when summarising news journalists want the key facts, while in other genres such as feature stories, reportage and interviews, the newsroom want a teaser effect. These genres have different structures and are not the same from the perspective of processing.

The media partners were particularly interested in how the summarization would function as teasers for paywalls (where other information could be used as well, e.g., information on time when the news is read). The discussions did not end with any conclusion as this function might be addressed in the future. Understanding what is popular in front of and behind paywalls would be interesting. Articles behind a paywall could be evaluated and rated to predict the popularity of similar articles depending on, for instance, the length and type of articles, or at what time they were published.

One group discussed how news archives could be useful when making news (see also Task 4.3). It is already normal to recycle parts of older news for background information. The archives could have a search function, which would be handy for finding similar keywords to the article piece that the journalist is writing. The search function could include quotes of politicians or categories for different topics with relevant keywords and cross lingual text capabilities (see also Task 4.3, related to archive analysis in terms of viewpoints and sentiment).

Task 4.3: Cross-lingual Identification of Viewpoints and Sentiment in News Reporting

(Task description: "For the news covering the same events and topics, the project will develop methods for detecting viewpoints and sentiments based on media sources.")

For media partners, regarding viewpoint, sentiment, and news reporting analysis five contexts could be useful.

- As a type of 'neutrality barometer' that could understand how the newspaper is positioned and tell if journalists are covering news from a "neutral" point of view or if they are very biased. Especially stance towards individuals is of interest (in relation to point 2). STY also mentioned journalist (author) embeddings models as potentially interesting, and fact-checking was mentioned as an important aspect.
- As information about how and how much journalists are covering certain topics or individuals (for example before elections, how are different candidates presented, including sentiment).
- For a comparison between countries on how same stories are covered and which are the interesting stories from neighbouring countries which should be reported (in relation to task T4.1).
- A tool that tells the journalist when a prominent individual or group (suddenly) changes opinions or stances on a particular subject. For this task, the analysis of quotes on topics over time would be useful.
- Tools for a diachronic analysis of archives was mentioned as particularly interesting for ExM (e.g., doping affair in sports).

WP5: Multilingual Text Generation

The objective is to design and develop news automation systems that are transferable across languages, transferable across domains, and are transparent in the natural language generation (NLG) process.

Task T5.1: Multilingual text generation from structured data



(Task description: "The task will build on the foundations laid out in task T2.3 and on-going work at UH-CS on language- and domain-independent NLG architectures. The task will start by selecting news topics and genres, together with the participating media partners, that will be used as the first example cases in development and prototyping.")

Task T5.2: Multilingual storytelling and dynamic content generation

(Task description: "We will get expert insights from experienced journalists in studying storytelling for various news domains in a multilingual setting. Using experts' insights on the topic, we will develop a novel method for automatically organising news articles based on the domain of the article, following the previous work of UH which studied ways of presenting news to be maximally informative to the assumed reader.")

The main partner in WP5 is the Finnish news agency STT, which have a company vision for automatisation, according to a representative.

Automatisation and robotics should be used in the newsrooms' workflow wherever it is possible and beneficial. STT wants to automate the workflows broadly: not only in one department or by focusing on a narrow subject. We will get there by focusing on projects which give us tools that are as general as possible.

The agency is particularly interested in news domains with data sources that allows local variation. The vision includes the aim to provide local media customers with locally customised automated news and/or datasets that focus individually on each customer's local area.

The news agency aims to automate all possible parts of the newsroom workflow. This will increase the value of STT's services in several ways:

- By letting the journalists focus on what the machines can't do: to ask questions, create context and sense the atmosphere.
- By creating speed and volume. By helping the news agency to follow subjects that can't be followed now.
- By allowing the news agency to provide better customised and localised news and data to it's customers.

An important question raised was if texts should be generated directly for the end customer (reader), or whether the generated texts should be considered tools and raw material for journalists and editors. The response leans towards the latter. As a tool, NLG can give both headline and whole story suggestions that journalists and editors in turn could choose to edit, use fully, or discard at their own volition. In the case of EMBEDDIA, texts produced by the system would be considered a starting point for editors and journalistic analysis, rather than an end product in itself. For evaluation purposes, the system needs to be designed so that the edits that journalists do can be tracked.

There were also discussions on whether the system could be developed to learn several different writing styles, be it the classic journalistic inverted pyramid with the most important items on top (Pöttker, 2003), a more chronological style, or a feature structure with drama; since different news require different ways of presenting information. How the system determines what is the most important and identifies the newest information automatically remains an open question to reflect on. Texts could be personalised to fit the level of the individual reader, either automatically or by letting readers choose which style they want for their news. Authorship is a crucial issue for publishers. How should automatically generated texts be labelled: should the computer be considered the author?

NLG systems are fully dependent on the supply of structured data. Much of the discussion therefore revolved around what type of data and news could be gathered and used. The list below shows (in no particular order) some types of news the media partners would be interested in automating:

Elections



- Hyperlocal news
- Sports
- Annual tax reports
- Incident reports by police
- Crime statistics
- Data of officials (for instance politicians' assets)
- Company board data
- Price news (food, gas)
- Changes in stock prices
- Significant changes in weather and the local environment
- Different types of events
- Public spending
- Betting data
- Unemployment data
- Urban planning
- Traffic info, for instance road delays
- Infrastructure and development plans
- Open public data, for instance the European Data Portal
- Real-estate market development

Certain topics fall under the umbrella term 'event based reporting', e.g., elections, annual tax reports, sports etc. while others are more broad, e.g., changes in weather, company board data, price news etc. These need to be categorised, as there are different applications for them. For example, with elections it is important to keep in mind that election data differs depending on country. While structured information is usually provided in the form of historic data in a tabular form, it is imperative when generating news that the data used is 'fresh'/'hot'. There are two types of applications for elections: comparison of previous results and real-time applications. For sport events, the historic and current data is usually in the same format and usually needs to be bought, in some instances it has to be scraped. Statistics in sports could further be used to produce reports on how recent games changed the standings of teams.

When trying to localize the news it was mentioned that previous texts could be used as data, e.g., text-to-text generation (or text-as-source). Even though outside the scope of the the EMBEDDIA project, this would be particularly useful when reporting on police and weather reports, or other data that is often already generated as short factual text statements. Moreover, the system could combine different sources of data from meteorological and other environmental institutes to generate a more creative or customized report, for instance on weather and road conditions in a local area. However, media partners do not posses structured data and it is critical that they be instructed on how to format their data. In addition to this, text-to-text generation is outside the scope of the work conducted within the EMBEDDIA project.

Keeping track of changes in data sources or alerting when something significant happens, passing the threshold set, was seen as very interesting for the media partners (see Plattner et al., 2016). For some news this is particularly important and interesting, e.g., company board data tracking (on public websites or sources) and changes in consumer pricing (market baskets). The weather in combination with sports is also an area of interest, e.g., how glaciers affect skiing, how snow or rain affects biking, or water temperature and swimming. Ideally, there could be a system in place to combine multiple, seemingly unrelated sources and combine them into one coherent report or overview.

Task T5.3: Creative language use for multilingual news and headline generation



(Task description: "We will make the generated texts more varied and colorful by generating creative expressions, especially in headlines. We will find similar terms and metaphors by finding analogous terms in different contexts using context-dependent embeddings.")

The workshop discussions concerning creative language showed that the needs regarding both news generation, and headline generation differ depending on the profile of the user. For instance, a newspaper that wants to position itself for different publics would want to incorporate more creativity in the texts than a news agency that aims to be more factual. The participants in the workshop agreed that clickbait in headlines should generally be avoided, although it was noted that some publications might prefer a more clickbait style. There is also an issue of fitting automatic headline generation into the specific content of the publisher. It was pointed out that it is vital to hold a continuous dialogue between tasks of "headlines as summary" (T5.3) and "news as summary" (T4.2).

4 Conclusions

The purpose of this report is to derive and prioritise user needs of the news media industry for the EMBEDDIA project. We analyse innovation challenges in news media and map the current state-ofthe art in media technology with regards to artificial intelligence tools in newsrooms. We present a project workshop where user needs were explored, applying action-research paired with user-oriented design as the methodology. The results from roundtable discussions in the workshop provide guidance for the development work in the project. This concluding part presents the ideas that were most prominently discussed in the workshop.

4.1 Recommendations

Below we list the three key recommendations from the workshop. However, it should be noted that during the progress of the project this preliminary guidance will continue to be updated based on evaluation what is important, preferable, and technically feasibile.

Recommendation 1: Comment management

In WP3, EMBEDDIA will focus on different aspects of user comments. For instance, the media partners expressed an interest in profiling the properties and opinions of their users, to compare and find both similarities and differences between profiles, as well as to map out how they react to different stories. Detecting comments that should be blocked or referred to human moderators is important. These could include many categories: abuse, hate speech, spam, self-promotion, use of the wrong language, trolling behaviour etc. Important components of this recommendation is;

- Automatic labelling with the reason for these blocking/moderation decisions.
- Detection of users who are particularly likely to generate comments needing blocking/moderation.
- Detection of positive/constructive comments (see T3.1 above).
- Detection of commenters' motivations and biases.

A remark from ExM representative indicates that adding these features to the EMBEDDIA Media Assistant would be important:

Definitely interested in all analysis and improvements we can find, as manual comments editing with a few moderators is not nearly sufficient enough to guarantee 'clean' comments section.

This remark was supported by comments from a STY representative:



We perform filtering unacceptable content and trolling detection manually and would like to automate it.

Recommendation 2: Detection of interesting news

In WP4, the media partners were keen on developing an alert system that detects breaking news or a similar system, which would identify interesting news to report on from the neighbouring countries. In addition, such a system could also be able to detect trending topics and keywords.

Here is an observation from an ExM representative:

I think that successfully detecting news (from different languages) that might interest our readers would be the most practical tool for us and also other news organisations. I think we should spend quite some time to discuss different criteria on what kind of news we are trying to detect this way (certain topics, popular news etc) and what kind of tools could be useful (automatic messaging, search etc).

In addition, many tasks related to news archive analysis were identified as relevant, as it is important to have background information available, and recycle part of old news when making new news. A few examples given were searching quotes said by politicians or general information on what happened during previous elections, and chronological analysis of archive related to specific topics (e.g., sport scandals).

Recommendation 3: Personalized news generation

In WP5, there was a keen interest in a NLG system that could generate texts styled for different genres, such as news pieces adhering to the traditional inverted pyramid, where is information presented in an order based on how important they are, but also, in chronologic order or a feature structure based on dramatic elements. Further, since NLG systems will work only with an input of structured data, news domains are dependent on access to such data. The Finnish news agency STT will focus on news domains and datasets that are of interest to their customers. Data management strategies based on editorial priorities will therefore be a necessity. The data issue is discussed more in detail in "D5.1: Datasets, benchmarks and evaluation metrics for multilingual text generation ". Who are the users? In the discussions, the consensus were that these text are generally not suited for publication as such but as raw material for journalists and editors who can refine outputs for different purposes. They can, for instance serve as alarms or updates on emerging stories.

4.2 Discussion

A limitation of this report is that we present preliminary conclusions after only nine months of working in the project. The mapping of state-of-the-art shows fast adoption and broad diversity of AI tools in news media in different parts of editorial work, from the production of content to distribution, analysis, personalisation etc. As this is an iterative process, we will continue to closely follow the tech development and update our knowledge. The discussion point related to technological development is that ideas conceived for EMBEDDIA in the project application stage need to be assessed again continuously during the three year research period. However, there are still only a few signs that AI technologies will be available in other languages than the major ones such as English, German, Spanish or Chinese, which means that the developed tools will fill a much-needed space.

An example of the iterative process is WP5 where EMBEDDIA researchers will attend a workshop on October 14, 2019 organised by STT for the company's customers. During the day, researchers will be able to discuss what these media companies see as potential applications for AI tools in the newsroom. STT sees a need to expand its business and EMBEDDIA could provide opportunities, for instance, users comment analytics on behalf of customers. STT representative notes in the survey: "Filtering could be interesting in the future if we were able to do that on behalf of our customers. It could be a service that we sell for newspapers." However, as another representative says: "[...] we



don't manage a web page etc. with user comments which we would monitor/moderate. At the moment we don't have an urgent need for user comment analysis."

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction machines: The simple economics of artificial intelligence. Boston, Massachusetts: Harvard Business Press.

Ali, O., Flaounas, I., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2010). Automating news content analysis: An application to gender bias and readability. Proceedings of the First Workshop on Applications of Pattern Analysis, 36-43.

Appelgren, E., & Nygren, G. (2014). Data journalism in Sweden: Introducing new methods and genres of journalism into "old" organizations. Digital Journalism, 2(3), 394-405.

Bastos, M. and Walker, S.T. (April 11, 2018). Facebook's data lockdown is a disaster for academic researchers. Available: https://phys.org/news/2018-04-facebook-lockdown-disaster-academic.html [Jun 12, 2018].

Beam, M. A. (2014). Automating the news: How personalized news recommender system design choices impact news reception. Communication Research, 41(8), 1019-1041.

Boczkowski, P. J. (2005). Digitizing the news: Innovation in online newspapers. Cambridge, MA: MIT Press.

Bosmajian, H. A., & Bosmajian, H. A. (1992). Metaphor and reason in judicial opinions. Carbondale, II: Southern Illinois University Press

Boyles, J. L., & Meyer, E. (2017). Newsrooms accommodate data-based news work. Newspaper Research Journal, 38(4), 428-438.

Brennen, J. S., Howard, P. N., & Nielsen, R. K. (2018). An industry-led debate: How UK media cover artificial intelligence. Reuters Institute for the Study of Journalism Fact Sheet.

Burns, E. (2015). How the New York Times uses predictive analytics algorithms. Available: <u>http://searchbusinessanalytics.techtarget.com/feature/How-The-New-York-Times-uses-predictive-analytics-algorithms</u> [29.12.2015]

Carlson, M. (2018). Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. New Media & Society, 20(5), 1755-1772.

Caswell, D., & Dörr, K. (2019). Automating complex news stories by capturing news events as data. Journalism Practice, 13(8), 951-955.

Caswell, D., Russell, F., & Adair, B. (2015). Editorial aspects of reporting into structured narratives. Proceedings of the 2015 Computation Journalism Symposium.

Chadwick, A. (2013). The hybrid media system: Politics and power. Oxford: Oxford University Press.

Chua, R. (2018). The cybernetic newsroom: Horses and cars. Retrieved from <u>https://www.reuters.com/article/rpb-cyber-idUSKCN1G00Z0</u> [12.08.2019]

Cooper, A., Reimann, R., & Dubberly, H. (2003). About face 2.0: The essentials of interaction design Indianapolis, In: John Wiley & Sons, Inc.

Cunningham, J. (2014). Academic discourse. In D. Coghland, & M. Brydon-Miller (Eds.), The SAGE Encyclopedia of Action Research (pp. 1-3). Los Angeles, CA: SAGE Publications.

Defrijna, S., Mathijsb, E., Gulinckc, H., & Lauwersa, L. (2008). Facilitating and evaluating farmer innovations towards more sustainable energy and material flows: Case-study in Flanders. Empowerment of the Rural Actors: A Renewal of Farming Systems Perspectives: 8th European IFSA Symposium, Clermont-Ferrand, France, 6-10 July 2008, 765-773.

Diakopoulos, N. (2019). Automating the news: How algorithms are rewriting the media. Cambridge, Massachusetts: Harvard University Press.

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. Digital Journalism, 4(6), 700-722.



Dupagne, M., & Chuan, C. (2019). Concluding comments: A few more points about artificial intelligence. Journalism & Mass Communication Quarterly, 1-23.

Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. Menlo Park, CA: Stanford Research Institute.

Ess, C. M. (2014). Editor's introduction: Innovations in the newsroom – and beyond. Journal of Media Innovations, 1(2), 1-9.

Grubenmann, S. (2016). Action research. Digital Journalism, 4(1), 160-176.

Guzman, A. L., & Lewis, S. C. (2019). Artificial intelligence and communication: A Human–Machine communication research agenda. New Media & Society.

Hibbitts, B. J. (1994). Making sense of metaphors: Visuality, aurality, and the reconfiguration of american legal discourse. Cardozo L.Rev., 16, 229.

Huovelin, J., Gross, O., Solin, O., Linden, K., Maisala, S. P. T., Oittinen, T., Toivonen, H., Niemi, J., Silfverberg, M. (2013). Software newsroom–an approach to automation of news search and editing. Journal of Print Media Technology Research. 2(3), 141-1556.

Inayatullah, S., Izgarjan, A., Kuusi, O., & Minkkinen, M. (2016). Metaphors in futures research. Futures. 84, 109-114.

Järventie-Thesleff, R., Moisander, J., & Villi, M. (2014). The strategic challenge of continuous change in multi-platform media organizations—A strategy-as-practice perspective. International Journal on Media Management, 16(3-4), 123-138.

Jiang, L., & Han, E.ModBot: Automatic comments moderation. (2019). Proceedings of the Computation + Journalism Symposium.

Karlsson, S. (2019). Structured data, not internet scraping, results in trustworthy robot-produced journalism. Retrieved from <u>https://www.inma.org/blogs/big-data-for-news-publishers/post.cfm/structured-data-not-internet-scraping-results-in-trustworthy-robot-produced-journalism</u> [26.08.2019]

Lehtisaari, K., Karppinen, K., Harjuniemi, T., Grönlund, M., Lindén, C., Nieminen, H., & Viljakainen, A. (2012). Media convergence and business models: Responses of Finnish daily newspapers. (No. 4/2012). Helsinki: Communication Research Centre CRC, University of Helsinki.

Leppänen, L., Munezero, M., Sirén-Heikel, S., Granroth-Wilding, M., & Toivonen, H. (2017). (2017). Finding and expressing news from structured data. Proceedings of the 21st International Academic Mindtrek Conference, 174-183.

Lewis, S. C. (2019). Artificial intelligence and journalism. Journalism & Mass Communication Quarterly, 1-23.

Liang, T., Lai, H., & Ku, Y. (2006). Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. Journal of Management Information Systems, 23(3), 45-70.

Linden, T. C., Tuulonen, H. E., Bäck, A., Diakopoulos, N., Granroth-Wilding, M., Haapanen, L., Leppänen, L., Melin, M., Moring, T., Munezero, M. D. (2019). News automation: The rewards, risks and realities of 'machine journalism'. Frankfurt. WAN-IFRA.

Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., Pomerville, S., Wudali, R. (2016). Reuters tracer: A large scale system of detecting & verifying realtime news events from Twitter. Proceedings of the 25th ACM International Conference on Information and Knowledge Management, 207-216.

Lokot, T., & Diakopoulos, N. (2016). News bots: Automating news and information dissemination on twitter. Digital Journalism, 4(6), 682-699.

Løvlie, A. S. (2016). Designing communication design. The Journal of Media Innovations, 3(2), 72-87.

Magnusson, M., Finnäs, J., & Wallentin, L. (2016). Finding the news lead in the data haystack: Automated local data journalism using crime data. Computation + Journalism Symposium.

Marconi, F., & Siegman, A. (2017). The future of augmented journalism: A guide for newsrooms in the age of smart machines. New York: AP Insights.



McDowell, W. S. (2011). The brand management crisis facing the business of journalism. The International Journal on Media Management, 13(1), 37-51.

Milosavljević, M., & Vobič, I. (2019, forthcoming). 'Our task is to demystify fears': Analysing newsroom management of automation in journalism. Journalism. 1-19.

Moeller, J. and Helberger, N., 2018. Beyond the filter bubble: Concepts, myths, evidence and issues for future debates. Amsterdam: University of Amsterdam.

Napoli, P. M. (1999). The marketplace of ideas metaphor in communications regulation. Journal of Communication, 49(4), 151-169.

Nyre, L. (2014). Media design method. The Journal of Media Innovations, 1(1), 86-109.

Nyre, L., Ribeiro, J., & Tessem, B. (2017). Business models for academic prototypes: A new approach to media innovation. The Journal of Media Innovations, 4(2), 4-19.

Perretti, M. (2019, May 21,). What journalists talk about when they talk about AI. Retrieved from <u>https://blogs.lse.ac.uk/polis/2019/05/21/what-journalists-talk-about-when-they-talk-about-ai/</u>[18.06.2019]

Plattner, T., Orel, D., & Steiner, O. (2016). Flexible data scraping, multi-language indexing, entity extraction and taxonomies: Tadam, a swiss tool to deal with huge amounts of unstructured data. Computation + Journalism Symposium.

Pöttker, H. (2003). News and its communicative quality: The inverted pyramid—when and why did it appear? Journalism Studies, 4(4), 501-511.

Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. New Haven and London: Yale University Press.

Rogers, S., Schwabish, J., & Bowers, D. (2017). Data journalism in 2017: The current state and challenges facing the field today. Retrieved from <u>https://newslab.withgoogle.com/assets/docs/data-journalism-in-2017.pdf</u> [16.08.2019]

Schein, E. H. (1987). The clinical perspective in fieldwork. Newbury Park, CA: Sage Publications, Inc.

Sirén-Heikel, S., Leppänen, L., Lindén, C., & Bäck, A. (2019). Unboxing news automation. Nordic Journal of Media Studies.

Stray, J. (2019). Making artificial intelligence work for investigative journalism. Digital Journalism, 1-22.

Stringer, E. T. (Ed.). (1996). Action research. A handbook for practitioners. Thousand Oaks, CA: SAGE Publications, Inc.

University of Helsinki/Reaktor. (2018). Elements of Al. Open access cource. Retrieved from https://course.elementsofai.com/

Veryzer, R. W., & Borja de Mozota, B. (2005). The impact of user-oriented design on new product development: An examination of fundamental relationships. Journal of Product Innovation Management, 22(2), 128-143.

Villi, M., Grönlund, M., Linden, C., Lehtisaari, K., Mierzejewska, B., Picard, R. G., & Röpnack, A. (2019). "They're a little bit squeezed in the middle": Strategic challenges for innovation in US metropolitan newspaper organisations. Journal of Media Business Studies, 1-18.

Vredenberg, K., Isensee, S., & Righi, C. (2001). User-centered design: An integrated approach. Upper Saddle River, NJ: Prentice Hall PTR.

Walker, S., Mercea, D. and Bastos, M.T. (2019). The Disinformation Landscape and the Lockdown of Social Platforms. Information, Communication and Society, 22(11), pp. 1531-1543.

Weber, W., Engebretsen, M., & Kennedy, H. (2018). Data stories: Rethinking journalistic storytelling in the context of data journalism. SComS Studies in Communication Sciences, 2018(1), 191-206.

Weghe, T. V. d. (2018). 10 things about AI every newsroom should know. Retrieved from <u>https://medium.com/jsk-class-of-2019/10-things-about-ai-every-newsroom-should-know-</u>19745dac8ad7 [09.08.2019]

Zachos, K., Apostolou, D., Paraskevopoulos, F., Ientsek, S., Maiden, N., Brown, A., & Mentzas, G. (2018). (2018). Creative information exploration in journalism. 9th International Conference on Information, Intelligence, Systems and Applications (IISA), 1-7.



Appendix 1: Participation in Computation + Journalism Symposium

Computation + Journalism Symposium 2019 <u>http://cplusj.org/</u> was held at the University of Miami, Florida on February 1-2, 2019. This is a small but very high class gathering of computer scientists and media people, both practitioners and researchers with a common interest in computational journalism. Lindén has previously attended these conferences at Stanford (2016) and Northestern (2017).

This year, C.G.Lindén from UH did not present a paper there, but he interviewed media managers for the Deliverable on user needs report (D6.3). The list includes names such as Lisa Gibbs and Stuart Myles (Associated Press), Titus Plattner, Didier Orel and Pietro Supino (Tamedia), Brian Hammand (New York Times) and Rick Hirsch (Miami Herald) but also researchers such as Nicholas Diakopoulos (Northwestern University), Irfan Essa (Georgia Institute of Technology), James Hamilton (Stanford University) and Laurence Dierickx (Université libre de Bruxelles) among many others.

Appendix 2: Questionnaire for survey

Automated Tasks for News Media

EMBEDDIA Project User Needs Survey

*Required

Name *

Your answer

Company *

Your answer

Country *

Your answer

Background / position in company *



User Comment Analysis *

	This isn't interesting	We do this already	We'd like to do this	We'd like to do it in multiple languages
FILTERING UNACCEPTABLE CONTENT: Removing comments which are not appropriate for posting on news portals, or that would be illegal to publish. Examples might be comments which contain hate speech; threatening material; homophobic, islamophobic, or agressive content.				
SENTIMENT ANNOTATION: Annotating user comments for sentiment (usually positive/negative/neutral, somtimes also with strength rating); either general sentiment, or stance towards a specific topic, person or organisation.				
EMOTION ANNOTATION: Annotating user comments with the strength and type of emotion expressed (e.g. anger, sadness, happiness, surprise).				
OPINION/ARGUMENT MINING: Analysing user comments with respect to arguments expressed, and their position for or against particular news topics (e.g. gun control, marijuana legalisation, Brexit).				
BIAS DETECTION: Detecting whether user comments are biased towards or against particular topics/events/people/organisations				
TROLLING DETECTION: Detecting comments whose purpose is to elicit extreme reactions from others (e.g. on political topics).				
CONTROVERSY DETECTION: Discovering topics which cause or are the subject of most controversy amongst users.				
INFLUENCE DETECTION: Discovering users whose comments are more influential in terms of information spread or effects on others.				
COMMUNITY DETECTION: Discovering groups of users from their commenting behaviour.				

User Comment Analysis: Any comments on the tasks above?

Your answer

User Comment Analysis: What other tool(s) would you like to see?



News Article Analysis *

News Alticle Allarysis				
	This isn't interesting	We do this already	We'd like to do this	We'd like to do it in multiple languages
LINKING RELATED NEWS: Linking news articles based on the topics/events they cover, so that similar past articles can be recommended to the reader.				
CATEGORISING NEWS: Categorising articles by topics (e.g. sport, foreign politics,) to help search, browsing.				
NEWS SPREADING: Analysing how articles/topics spread (e.g. who first publishes it, who takes it from you).				
IDENTIFYING SPECIFIC GENRE/TYPE: For example, identifying protest news (are there specific types/categories of particular interest?)				
SUMMARISATION AND VISUALISATION: Presenting content as e.g. interactive graphs of keywords, lists of important words/phrases or key text excerpts, lists of main events etc; visual aids (colours) to identify different sources, time points, languages, etc.				
NEWS ARCHIVE ANALYSIS: Analyzing content in a news archive, identifying main topics, important events				
TOPIC ANALYSIS ACROSS TIME: For a selected topic, identifying changes in content, focus, keywords, subtopics over time.				
SENTIMENT ANNOTATION: Identifying where positive or negative tone or stance is expressed (perhaps towards a particular event/topic/person/organisation)				
OPINION DETECTION: Which sentences of a news article are neutral and which contain opinions?				
VIEWPOINT/BIAS DETECTION: Does a news article/news source present an event/topic/personality with a particular stance, position or does it express bias towards/against a particular topic or stance? Which are the words characteristic for a specific viewpoint?				
DETECTING IMPORTANT NEWS: Rating news articles/topics for importance, perhaps with respect to different countries/audiences.				

News Article Analysis: Any comments on the tasks above?

Your answer

News Article Analysis: What other tool(s) would you like to see?



News Generation *

	This isn't interesting	We do this already	We'd like to do this	We'd like to do it in multiple languages
LEAD GENERATION: identifying potential news stories hidden in data. Ignore "breaking" news such as a sudden natural disaster in this answer.				
GENERATION OR REUSE OF TEXT FRAGMENTS: for e.g. inclusion in "fact boxes" or to act as building blocks for a human written news story				
AUTOMATIC PRODUCTION OF NEWS ARTICLES: producing complete or near- complete news articles (or reports) with no or very little human intervention. Please describe the type/domain of story produced and the type of data used in the comment section				
LINGUISTIC AND CREATIVITY AIDS: proposing novel metaphors, similes etc.				
HEADLINE GENERATION: fully or partly automated generation of either ready headlines or ideas for headlines				

News Generation: Any comments on the tasks above?