Intelligent Automation in Journalism

Are newsrooms ready to let machines write our news?

Sampo Sauri

Master's Thesis
Media Management
2022

DEGREE THESIS	
Arcada	
Degree Programme:	Media Management
Identification number:	8770
Author:	Sampo Sauri
Title:	Intelligent Automation in Journalism Are newsrooms ready to let machines write our news?
Supervisor (Arcada):	Mats Nylund
Commissioned by:	–

Abstract:

The advancements in readily available structured data, artificial intelligence and automation tools have led to newsrooms exploring the possibility of creating news articles automatically. Sports, finance, politics and weather are all fields of journalism where an abundance of available data makes it possible to automate the process from beginning to end, using natural language generation (NLG) techniques to convert the data to human-readable articles. The difference between models based on templating and models based on machine learning algorithms is explained. The thesis is based on literature and media reports and is complemented by expert interviews from Finnish public broadcaster Yle and news agency STT. The aim of the study is to describe the techniques used to automate news production, explore how the production of news is generally automated, and to further build on these findings by interviews, and make visible some ethical questions related to machine-generated news.

Keywords:	journalism, news automation, machine learning, artificial intelligence
Number of pages:	57 + 3 appendices
Language:	English
Date of acceptance:	16.6.2022

EXAMENSARBETE	
Arcada	
Utbildningsprogram:	Media Management
Identifikationsnummer:	8770
Författare:	Sampo Sauri
Arbetets namn:	Intelligent Automation in Journalism Are newsrooms ready to let machines write our news?
Handledare (Arcada):	Mats Nylund
Uppdragsgivare:	–

Sammandrag**:**

Tillgången till strukturerade data samt utvecklingen av artificiell intelligens och verktygen för automatisering har lett till det, att nyhetsredaktioner utreder möjligheterna till automatisk produktion av nyheter. Sport, ekonomi, politik och väder är alla grenar av journalism där mängden användbara data möjliggör automatiseringen av hela processen från början till slut med hjälp av NLG-tekniker. I slutarbetet beskrivs hur schablonbaserade modeller skiljer sig från modeller baserade på maskininlärning. Detta arbete baserar sig på litteratur och rapporter från mediabranschen och kompletteras av intervjuer med experter från Yle och notisbyrån STT. Avsikten med arbetet är att beskriva teknikerna som används för att automatisera nyhetsproduktionen, utreda hur den nyhetsproduktionen generellt automatiseras, och bygga på dessa med hjälp av intervjuer, samt klargöra kring vissa etiska frågor som maskingenererade nyheter hämtar med sig.

Nyckelord:	journalism, nyhetsautomatisering, maskininlärning, artificiell intelligens
Sidantal:	57 + 3 bilagor
Språk:	engelska
Datum för godkännande:	16.6.2022

OPINNÄYTE	
Arcada	
Koulutusohjelma:	Media Management
Tunnistenumero:	8770
Tekijä:	Sampo Sauri
Työn nimi:	Intelligent Automation in Journalism Are newsrooms ready to let machines write our news?
Työn ohjaaja (Arcada):	Mats Nylund
Toimeksiantaja:	–

Tiivistelmä:

Rakenteisen datan saatavuus sekä keinoälyn ja automatisointityökalujen kehitys ovat johtaneet siihen, että uutistoimitukset selvittävät mahdollisuuksia uutisten automatisoituun tuottamiseen. Urheilu, talous, politiikka ja sää ovat kaikki journalismin lajeja, joissa tarjolla olevan tiedon määrä mahdollistaa prosessin automatisoinnin alusta loppuun NLG-tekniikoita käyttämällä. Työssä selitetään myös millä tavoin sapluunapohjaiset mallit ja koneoppimispohjaiset mallit eroavat toisistaan. Päättötyö pohjautuu kirjallisuuteen ja media-alan raportteihin ja täydentyy yleisradioyhtiö Ylen ja uutistoimisto STT:n asiantuntijoiden haastatteluilla. Päättötyön tarkoituksena on kuvailla automaattisen uutistuotannon tekniikoita, tutkia kuinka uutistuotantoa yleisesti ottaen automatisoidaan ja syventää löydöksiä haastattelujen avulla sekä tuoda näkyväksi koneiden tuottamien uutisten eettisiä kysymyksiä.

Avainsanat:	journalismi, automaattinen uutistuotanto, koneoppiminen, keinoäly
Sivumäärä:	57 + 3 liitettä
Kieli:	englanti
Hyväksymispäivämäärä:	26.6.2022

CONTENTS

FIGURES

FOREWORD

The idea for this thesis came to me from wanting to better understand the buzz words of the past decade – artificial intelligence, machine learning, algorithms – and by having been intrigued by the possibility of AI in the form of computer vision and image generation, such as deepfakes, filters that turn photos into paintings, and FaceApp that turns a frown into a smile, to name a few. I wished to learn if and how these methods were used in written journalism, i.e. printed and online news. In addition, I wanted to contrast this by examining how my native Finland, a Nordic country with a traditionally very high newspaper readership – 92% of the population reads print or online newspapers weekly (Media Audit Finland, 2021) – compares with the rest of the world in with regard to adoption of this technology.

Artificial intelligence as a term means different things to different people, and suffers from hype, which is reason enough to avoid having it in the title of this thesis. The terminology in this field is overlapping and has yet to find a universal standard. The term I've found to best capture both low-level automations based on templating, and high-level automations using machine learning algorithms or natural language processing that are used to automatically produce news articles at scale, is Intelligent Automation (IA). Here, it is applied to automatic generation of content in journalism, but the term has its roots in the automation of business processes (Cognizant, 2022). Intelligent Automation as a term has been used in a journalism context as well (Newman, 2022, p. 35).

This thesis is divided in five distinct chapters. First, the reader is introduced to media's relationship with technology, what is meant – and not meant – by artificial intelligence, and definitions for what core concepts mean in the context of this thesis.

Second, I present the methods used for gathering data and conducting interviews, and explain the rationale behind said choices. The people interviewed and the organizations they represent are introduced to the reader.

Third, I go into some detail to explain how article creation is automated based on literature and theory. I explain what the different types of automation are, and what kind of data they require, and present real-world examples of solutions that have been

implemented in newsrooms internationally. In this chapter, I also examine the impact news automation has on the industry, and what kind of ethical questions it raises.

Fourth, I let the people interviewed shine, and explore the experiences they have had in Finnish news automation.

In the fifth and final chapter, I attempt to sum up my findings in an easy-to-grasp package and provide an estimate on where this Intelligent Automation in news production is headed next.

The field of automation, computational journalism and artificial intelligence is broad, to put it mildly. The terminology is partly overlapping and has yet to find a standardized, universally applicable form. Also, these new tools affect many – if not all – parts of the media industry, so deliberate choices to limit the scope of this thesis have been made. The emphasis is on the *production* of news instead of the *publication* and its tangents: recommendation engines and personalization. Paradoxically, my chosen area is the least important use of AI in journalism according to news leaders around the world – but perhaps the most future-focused of all (Newman, 2022, p. 35).

My hope is that after reading this thesis, you will have a firm grasp on how newsrooms are using automation today, and how Intelligent Automation will help them create more meaningful journalism for their audience.

1 INTRODUCTION

> *There is no universal definition of artificial intelligence. For computer scientists, AI might look like algorithms capable of thinking like humans. For bioengineers, it might mean growing brain cells in a laboratory. But how should journalists think about AI? One way of thinking about AI in news organizations is in terms of the interaction between humans and machines and the journalistic results of that collaboration.*
>
> (Marconi, 2019, p. 55)

Journalists have always used different tools to write and convey the news – from the pen and paper to the typewriter and the word processor, from film cameras to DSLR's (digital single-lens reflex cameras) and smartphones, from the printing press to offset printing and online publication, just to name a few advances in technology pertaining to news. To put it literally, according to Merriam-Webster, the word *media* (the plural form of *medium*) means "a channel or system of communication, information, or entertainment". Media and technology have always been intertwined: as technology evolves, it is taken into use in the media. Automation and artificial intelligence are no exceptions.

1.1 The Link Between Media and Technology

In 1964, Marshall McLuhan famously wrote "the medium is the message" (McLuhan, 1964, p. 7), but it is argued that what he meant is not that the choice of technology to convey the message is important, but what kind of changes and noticeable social effects the use of it brings with it. (Federman, 2004). Regarding artificial intelligence producing the news, in effect relinquishing a part of the journalistic process from a human to a machine, the potential for such change is huge. Like it or not, this technology, like other technological advancements before it, will change the way news are produced, published, and consumed.

The big question is, will these new tools be used to assist journalists in producing better news, automating tedious tasks such as data collection, to free up journalists' time to do more creative and meaningful work, such as investigative journalism, or are they seen merely as a means to cut costs at the newsroom.

In this thesis, the term "computational journalism" (CJ) is used as an umbrella to describe news produced by these newfangled methods. Nicholas Diakopoulos defines computational journalism as "Finding and telling news stories, with, by, or about algorithms" (Diakopoulos, 2016). Prior to the general, nowadays ubiquitous, availability of computers in the newsroom and the onset of the more sophisticated tools this thesis focuses on, the term computer-assisted reporting (CAR) was used to describe journalism that was done by using computers – things that nowadays are taken for granted, such as word processors, spreadsheets and databases (Salmela, 2021, p. 5). According to Diakopoulos' (2016) interpretations of research in the field, computational journalism differs from CAR and data-driven journalism (DDJ) by being "rooted in applications of automation to information", thus being the appropriate term in the scope of this thesis – the automation of the production of news stories.

When viewed from a perspective of decades, this progression is completely logical: From computers assisting, to computers providing the data in an easily usable (and reusable) form, to computers going even further to make news.

Noam Lemelshtrich Latar divides the field of automated journalism into two distinct pillars: "The computer software that automatically extracts new knowledge from huge data silos, and algorithms that automatically convert these insights and knowledge into readable stories without human involvement." (Latar, 2019, p. 29)

The first pillar encompasses software that can analyze big amounts of data using different statistical models to find outliers, deviations from a pattern, or other interesting insights. A human journalist must evaluate the results and deem if the findings are newsworthy. This technique can also be called data analytics and is in no way exclusive to the field of journalism. Before the data can be analyzed, though, it must be made available: newsrooms traditionally use many different tools for data collection, including programmatically scraping the web to extract data, to public datasets, and in many cases the outlet's own archives. In these cases, it is important that the data is machine-readable, i.e. that a computer can extract the data and input it into a spreadsheet or database for further analysis.

Latar's second pillar describes the actual production of a news article by computer algorithms without human intervention (Latar, 2019, p. 29). This pillar forms the basis for the two completely different methods this thesis goes into lengths to describe and differentiate: on one hand there are the story templates written by human journalists, where depending on the data available, the algorithm fills in the blanks and churns out articles where the written text itself is constructed according to a predefined set of rules. On the other, texts wholly written by a computer model that has been trained by a machine learning algorithm using vast amounts of data, such as news archives, to effectively emulate a real human journalists' rich and diverse language.

It is on this second pillar that this thesis sets its focus: the automation of news production using either template-based processes (in very simple terms not unlike a choose-your-own-adventure -book) or processes based on machine learning.

As always with the arrival of new technology, there is the hope that it will change the world for the better and make work easier or at least free up journalists' time for more meaningful tasks. On the other hand, some fear this newfangled technology is instead rendering human journalists obsolete, or at least make them face strong competition from robot journalists. (Latar, 2019, p. 29)

Already in the 1980's, Shoshana Zuboff – one of the original pioneers of the information age – studied the computerization of factories and offices, and crafted the three so-called Zuboff's laws:

> *1. Everything that can be automated will be automated.*
>
> *2. Everything that can be informated will be informated.*
>
> *3. In the absence of countervailing restrictions and sanctions, every digital application that can be used for surveillance and control will be used for surveillance and control, irrespective of its originating intention.*
>
> (Zuboff, 2013)

Of these laws, the first and second – in reversed order – find their tangents also in Latar's pillars. Zuboff's third law also applies to the media industry, but it is not directly applicable to automated news, and will not be explored further in this thesis as such.

Through digitalization, the amount of data available is skyrocketing while simultaneously being increasingly quantifiable and readable by automated systems. Zuboff coined the word *informate* as meaning the added information content automated tasks bring with them. While Zuboff studied the future of work when she wrote this, the amount of data collected and the dynamics that can follow from the increased information content is relevant in other fields as well.

> *As long as the technology is treated narrowly in its automating function, it perpetuates the logic of the industrial machine that, over the course of this century, has made it possible to rationalize work while decreasing the dependence on human skills. However, when the technology also informates the processes to which it is applied, it increases the explicit information content of tasks and sets into motion a series of dynamics that will ultimately reconfigure the nature of work and the social relationships that organize productive activity.*
>
> (Zuboff, 1988, p. 10–11)

This increased information content could today be called big data, and the field of utilizing it and mapping the complex dynamics involved could be data analytics or just data science – a pivotal part of Latar's first pillar.

When Zuboff's first law meets Latar's second pillar, one can draw the conclusion that every aspect of a journalist's work that can be automated, will be. The limit to what journalistic tasks will be automated is technological. Logically, routine tasks that do not need a high level of creative intelligence will be the first to be automated, while the more creative tasks still need a human touch.

The relationship of Latar's pillars and Zuboff's first two laws and the process of data to knowledge to articles to complex dynamics are depicted in Figure 1.

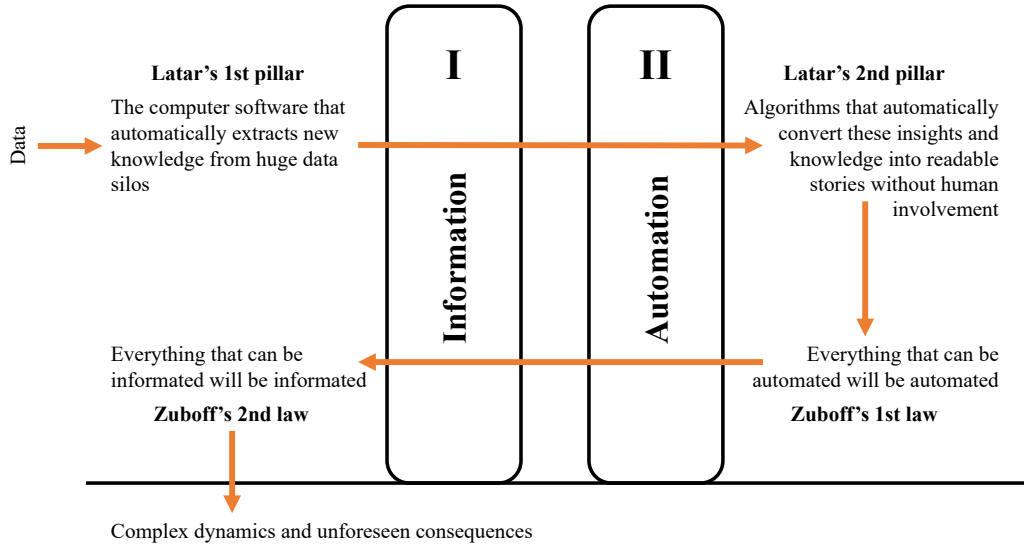

Figure 1. Relationship between Latar's pillars and Zuboff's first two laws (Latar, 2019, and Zuboff, 2013, compiled by the author)

While automated journalism brings about changes in the newsroom, the possibility of automatically analyzing vast amounts of data and publishing stories based on them may bring about complex dynamics that are not only changing the way journalists work, but also have effects on society at large. The algorithms may allow monitoring of large-scale phenomena, but only understand them to the extent programmed. AI cannot be left to monitor itself, either, as it will not alert society about its own pitfalls, or the effects of more pervasive introduction of AI into human lives. Human journalists must be aware of technological developments. (Latar, 2019, p. 25).

It is important to note, here, that this automation is not necessarily detrimental to the journalist's work and can more often than not be seen as tools to be used instead of rivals to compete against. Automated journalism gives the opportunity to cater to new audiences, faster, and with less errors and bias – if used properly. (Latar, 2019, p. 29). Leveraging machine learning algorithms has the potential to improve these capabilities even further.

1.2 Definition of Artificial Intelligence

Artificial intelligence (AI) does in fact not currently have a universally accepted definition, despite an increased interest in the topic. Many definitions of artificial intelligence "refer to machines that behave like humans or are capable of actions that require intelligence", but as it is also difficult to define and measure human intelligence it is hard to come by an objective or specific definition. (Samoili, et al., 2020, p. 6). What, then, constitutes as AI generally, and what is meant by AI in the scope of this thesis?

The High-Level Expert Group on Artifical Intelligence (2019) appointed by the European Commission, defines AI as follows:

> *Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*
>
> *As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).*

The High-Level Expert Group's definition of AI is technical and comprehensive, and as such is not instantly approachable. Based on the above definition that "AI systems can either use symbolic rules or learn a numeric model" it is apparent that both template-based news production (using symbolic AI, also coined "good old-fashioned AI" or GOFAI (Haugeland, 1985, p. 118)) and machine learning -based news production using

neural networks could potentially fit under the umbrella term of artificial intelligence. However, the basic templates such as “simple code that extracts numbers from a database” (Graefe, 2016) are not considered sophisticated enough to warrant being called “intelligence” instead of just “automation”.

Template-based content generators are intelligent in the same good old-fashioned way that for instance the chess machine Deep Blue that beat grand master Garry Kasparov in 1997 is. They can, according to human-crafted rules, calculate the most effective move, or select the most appropriate template snippets and populate placeholder fields with numbers or names in order to generate an article that looks like it could have been written by a human. While this kind of GOFAI might be useful and effective, it is not especially creative. (Hyppönen, 2020).

This is where machine learning differs from template-based automation. A model built by analyzing vast amounts of pre-existing material can be provided with parameters and given a task to create something new. However, as the current AI algorithms have such a limited understanding of how humans communicate, automation cannot be creative or have the necessary context to generate new ideas, metaphors or humor, as they “lack the the human capability to make connections not previously experienced”. (Latar, 2019, p. 25).

It is also worth noting here that AI can also be divided into narrow (or weak) AI that can only perform specific, limited tasks, and general (or strong) AI that can perform most activities that humans do. (AI HLEG, 2019). The applications described herein are decidedly narrow, only pertaining to a specific task – the creating of a news article or some sub-task thereof. Human journalists will not find themselves working alongside walking, talking robot colleagues that can do what they do – at least not in the foreseeable future. But machines are taking on more and more of the tasks usually performed by humans. The Future Today Institute’s 2022 Tech Report lifted computer-directed reporting to its list of news & information related trends: “Computer-directed reporting applies natural language processing (NLP) algorithms and artificial intelligence to automate many common tasks like curating a homepage and writing basic news stories.” (Future Today Institute, 2022 (a), p. 8).

1.3 Research Questions

The aim of this thesis is to explain how journalism is affected by the onset of new, automated tools that can write news without a human's input. In order to do so, it is important to first describe the techniques with which this automatic news generation is performed, i.e. the template-based and machine learning -based approaches.

The primary research question for this study is:

- How do news organizations use automation to write the news?

The secondary research questions, aimed to support the primary as well as answer possible follow-up-questions, are as follows:

- What are the reasons for news organizations to adopt automation techniques?
- What effects does automation have on news organizations?
- What are some of the implications that automation may bring about in society?

1.4 Core Concepts

The usage of computer automation in the newsroom is relatively new, so it is best to define how some core concepts are used in the scope of this thesis. The terminology is still in its infancy, and in many cases the meaning of different expressions overlaps each other. (Beckett, 2019, pp. 15–19).

automated journalism	"the use of automation in the production of written news content" (Diakopoulos, 2015)
automatic narration	the creation of readable stories by algorithms leveraging data from existing datasets
artificial intelligence (AI)	see chapter Definition of Artificial Intelligence (p. 13). It is used by different people to mean different things. (Beckett, 2019, p. 92).
computational journalism (CJ)	"the combination of algorithms, data, and knowledge from the social sciences to supplement

	the accountability function of journalism" (Hamilton & Turner, 2009 p. 2)
computer-aided reporting (CAR)	as reporting is seldom done without computers, "what used to be called CAR has mutated into a variety of different categories" (Kjellman, 2021, p. 16)
datafication	the creation of measurable data from observations (Diakopoulos, 2019, p. 117)
data journalism	"the umbrella term most commonly used to describe journalistic practices that rely on analysing and presenting data" (Kjellman, 2021, p. 16)
intelligent automation (IA)	combines "robotic process automation with advanced technologies such as artificial intelligence, analytics, optical character recognition, intelligent character recognition and process mining to create end-to-end business processes that think, learn and adapt on their own" (Cognizant, 2022)
machine learning (ML)	an application of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed
natural language generation (NLG)	natural language generation enables the automation of repetitive tasks like writing news articles that follow a well-defined structure." (Marconi, 2020, p. 60)

natural language processing (NLP)	the creation of human-understandable written content using machine learning algorithms trained on human-produced texts
neural network	"a program or system which is modelled on the human brain and is designed to imitate the brain's method of functioning, particularly the process of learning" (Collins Dictionary)
robot journalism	also dubbed **robo-journalism**. See **automated journalism**.
template	a human-written article piece, with blanks left to be filled from a dataset by an automated process, that when combined create an article

2 METHODS

This chapter describes the different methods used for gathering data for this thesis, as well as explains the rationale behind choosing that method. The methods used are a literature review of recent international publications, coupled with interviews with industry professionals to get an angle on how the topic is viewed by key news organizations in Finland.

2.1 Literature

This thesis leans heavily on recent literature on this topic. The two most prominent sources are *Automating the News* (2019) by Nicholas Diakopoulos and *Newsmakers* (2020) by Francesco Marconi. In addition to these, the thesis references several media industry reports and the book *Robot Journalism* (2018) by Noam Lemelshrich Latar.

2.2 Interviews

For the data gathering from local experts at Finnish national broadcaster Yle and Finnish News Agency STT, respectively, a semi-structured interview was conducted.

A semi-structured or focused interview "is defined as an interview with the purpose of obtaining descriptions of the life world of the interviewee in order to interpret the meaning of the described phenomena" (Kvale & Brinkmann, 2008, p. 3), which allows the interviewer the possibility of focusing on certain aspects of the interviewee's answers as well as to ask clarification when needed.

The interviewees were selected on basis of their expertise in the field. Jarkko Ryynänen is Project Manager at Yle Newslab, developing in-house solutions for the future of journalism. Yle in turn is the Finnish taxpayer-funded public service broadcaster, with a yearly turnaround of approx. 500 million euro. (Yle, 2022). Ryynänen's team is behind the Voitto robot, which started writing autonomous sports news in late 2015.

Salla Salmela is Producer (Robotics Projects) at STT. Her employer is a national news provider co-owned by 30 media companies. Its majority owner is Sanoma Group, publisher of Finland's largest daily newspaper Helsingin Sanomat, with a 75,42 per cent

share. Even though Yle has its own news service, it is a minority owner and has renewed its subscription to STT after a prolonged hiatus. (Ala-Fossi, et al, 2021; STT, 2022).

STT has been involved in different robotics and automation projects since 2017, including the Google-funded Digital News Initiative project Scoopmatic, which wrote sports news using machine learning, developed together with University of Turku researchers. (Paikkala, 2020).

The interview questions were pre-written, but the semi-structured interview method allows for the possibility of asking pertinent questions and to some effect steer the topic in the direction that is most useful for the research question at hand. The battery of questions was the same for both interviews, although they were deviated from to get the most relevant answers for the thesis.

> *Compared to structured interviews, semi-structured interviews can make better use of the knowledge-producing potentials of dialogues by allowing much more leeway for following up on whatever angles are deemed important by the interviewee. Semi-structured interviews also give the interviewer a greater chance of becoming visible as a knowledge-producing participant in the process itself, rather than hiding behind a preset interview guide. And, compared to unstructured interviews, the interviewer has a greater saying in focusing the conversation on issues that he or she deems important in relation to the research project.*
>
> (Brinkmann, 2013, p. 21)

Because of the ongoing covid-19 pandemic at the time of writing this thesis, the interviews were conducted using Microsoft Teams video conferencing software, with cameras on for both the interviewer and interviewees during the sessions in November and December 2021, respectively. The interviews were recorded and transcribed, and transcripts are attached to this thesis as Appendix I & II.

The interview with Jarkko Ryynänen was conducted in English, while the interview with Salla Salmela was conducted in Finnish. The latter transcript is attached in the original language, and the citations provided in the thesis are translations by the author.

3 THEORY

> *Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.*
>
> (Humby, 2006)

For news, or any content for that matter, to be produced automatically, the availability of data is the most limiting factor for success. Data is needed to drive the process. In this sense, Clive Humby, who first coined the popular expression "data is the new oil" in 2006, was right. In this case, data is the fuel for content creation, but it only becomes valuable through a process of analysis and refining.

However, these oil fields must be found to be utilized. Not all knowledge is created equal, as data from some fields are more easily collected than others, and not all sources of data are of the same quality. The quality of the raw material affects the quality of the end product.

> *Datafication—the process of creating data from observations of the world—becomes a stricture that holds back more widespread use of automation simply because aspects of the world that aren't digitized and represented as data cannot be algorithmically manipulated into content. The quality, breadth, and richness of available data all impact whether the automated content turns out compelling or bland.*
>
> (Diakopoulos, p. 117)

Automating the journalistic process does not work equally well in all fields of journalism because of the amount of readily available data that can be used to drive the process. Therefore, most efforts are currently focused on journalistic genres where there is a wealth of available statistics, such as sports and business. Journalists can produce a set of templates, and computers can fill in the gaps. (Galily, 2018, p. 106)

Like oil, the availability of data is also a way to set yourself apart from the competition. If a news outlet has its own source of data, or exclusive contracts with data providers, it

means they can provide exclusive content. (Diakopoulos, 2019, p. 117). As a journalist protects and safeguards their sources, it also makes sense for outlets to safeguard exclusive access to sources of structured data.

In recent years, many countries have passed legislation that mandates government-produced data to be made publicly available. For instance, the European Commission has passed a directive to harmonize data publication throughout the European Union (European Commission, 2022). Other countries have comparable laws, and there are also public, open, machine-readable data sources made available by other types of organizations. Data that is made available through API's or searchable databases or even downloadable spreadsheets are much easier for a computer to read than, say, Word documents or scanned pdf images, not to speak of paper copies in an archival cabinet. Investigative journalists' possibilities of uncovering wrongdoings are made significantly harder when data is closed behind technical barriers. In many countries there are also laws that give access to official-held information, such as the freedom-of-information-act (FOIA) in the US or similar requests to officials to get access to information that is public but not published.

The availability of high-quality data is needed for template-based journalism where each produced news article requires data in the same format for each article – for example what the name of the opposing teams are, how the game ended, the names of the players that scored a goal, et cetera – but it is especially true for articles that are being used as training material for a machine learning algorithm.

In terms of data and data quality, the five V's of a data from a given source – volume, velocity, variety, value and veracity – need to be considered. Volume means how much data is available and velocity how quickly more data is added. For automated journalism to be viable, data must either have volume or velocity or both. Variety refers to the type of data and its complexity, and value in this context that the data is newsworthy. The last one, veracity, is the most important. None of the other V's matter if the veracity – i.e. accuracy or trustworthiness – of the data cannot be guaranteed. Automated news, like all news, must be true. (Lindén, et al, 2019).

With high-quality data sources, one of the main gains is the massive potential for scale. Once a computer algorithm has been properly set up, it can generate hundreds or thousands of variations of a story, and provide content for ever smaller or niche audiences. (Diakopoulos, 2019, pp. 109–110).

Although automated news are designed to write news automatically, in many cases humans oversee the process. At Bloomberg, human intelligence is guiding the automation process, double-checking if there are errors. Journalists are needed to tell machines what to do, so the process is not completely automated. (Broussard & Lewis, 2019)

3.1 Types of Automation

> *Artificial intelligence technologies such as Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and Natural Language Generation (NLG) have become more embedded in every aspect of publishers' businesses over the last few years. Indeed, these can no longer be regarded as 'next generation' technologies but are fast becoming a core part of a modern news operation at every level – from newsgathering and production right through to distribution.*
>
> (Newman, 2022)

Most articles written by robots today are, in fact, not produced through any mysterious neural network, but instead automated through a multitude of *if ... then* statements that parse the provided data and generate an end result.

Automatically generating news articles is not the only way automated processes can aid in producing the news. Below, these are divided into three categories: templates, machine learning, and virtual assistants & article curation. The first two are different paradigms altogether, and in the third category are automated tools that help the journalist, not something that produces articles themselves.

3.1.1 Articles Generated by Templates

As mentioned, most automated journalism is in fact based on templates and programmatical *if ... then* statements. Two rudimentary examples of such a template can be found below:

```
Home sales in [town] measured [home sales this
year] this year, [a decline / an increase / staying
flat] compared to [home sales last year] sales rec-
orded last year.
```

and

```
[Team name] scored [adjective] [number of points]
in [quarter], as [player] led the way with [fre-
quency of scores] [types of scores]
```

(Marconi, 2019, pp. 60 & 82)

By substituting the word in brackets with actual data from a source, a real sentence that can be used as part of a news story is formed. In effect, coupled with trustworthy data sources that can be read programmatically, churning out articles can be automated to a high degree. The issue that arises with templates is that they need to be made and updated by IT-savvy journalists or engineering staff, otherwise their usefulness dwindles as the templates become outdated. Using technological solutions demands technical personnel to make sure the automation systems are kept up and running.

> *Lack of technical resources that are able to focus on simplification of production and distribution. Product, design and Engineering teams are dwindling in newsrooms and those who are growing are being treated as more of a service department than a strategic necessity. News organizations also need to start caring more about what their audience thinks than what their industry peers think. My hope is that we can disrupt ourselves before something else comes along and disrupts journalism for us.*
>
> (Future Today Institute, 2022 (b)).

In practice, templating software tends to be somewhat more complicated than the above examples. For instance, Figure 2 shows Yle's Voitto robot's article generation process as a flowchart.

In the flowchart, the article writing process is first started, then the app scours the score database for game goals and other data such as times of events, penalties, names of players and the like, which are then parsed, and suitable templates are chosen based on predetermined conditions. The robot then selects an applicable headline and image for the article, as well as a list of events or a table of statistics, and then puts all these bits and pieces one after the other. Lastly, it exports the article in JSON format that can then be fed into the publishing system, or copied as text for instance for further editing.

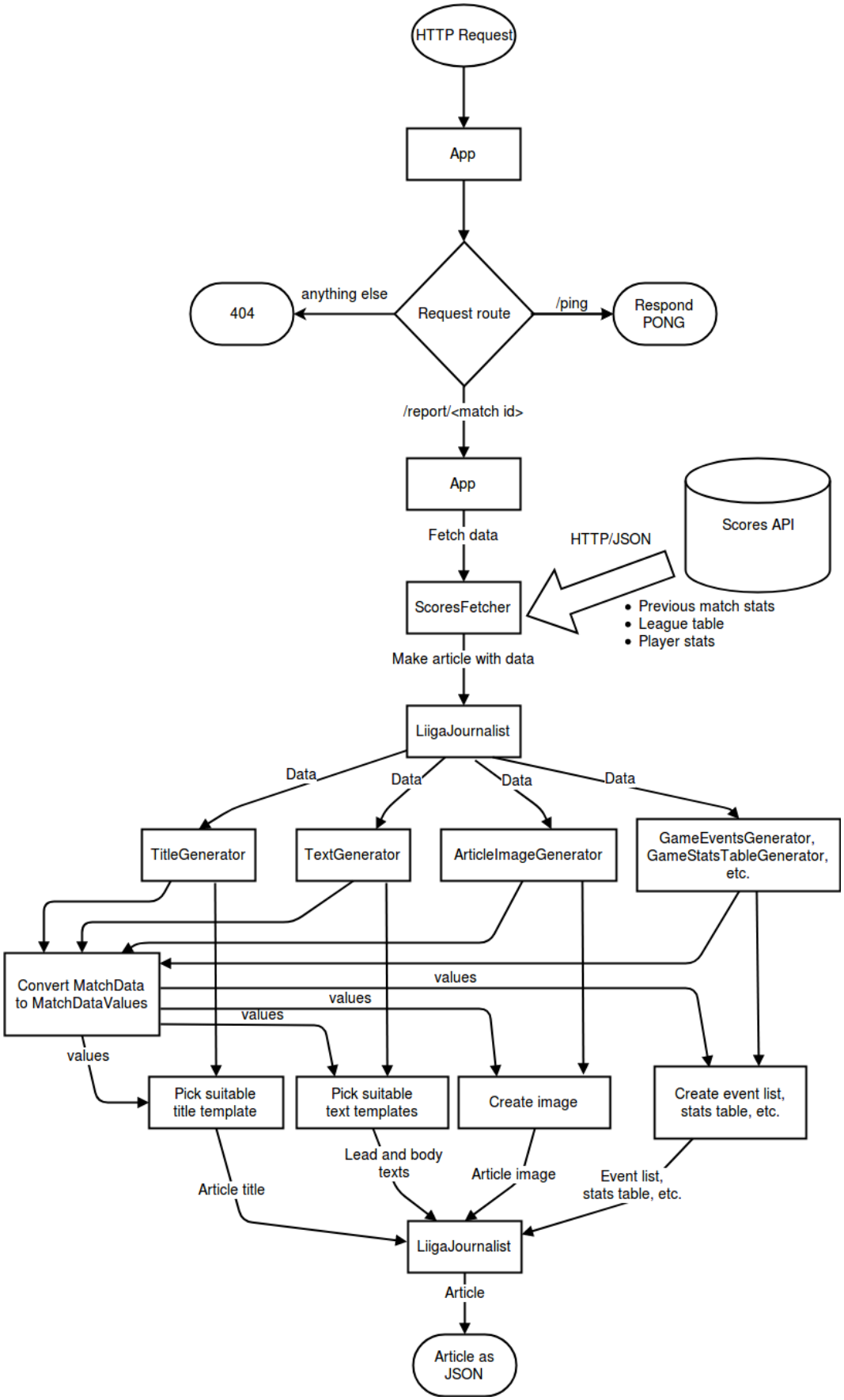

Figure 2. Flowchart of Yle's Voitto robot's article generation process (Yle, 2017)

3.1.2 Articles Generated by Machine Learning

In 2020, The Guardian newspaper published an opinion piece written completely by artificial intelligence, although it was edited before publication, as is custom for all op-eds. The AI produced eight essays, of which The Guardian selected the best bits and combined them into one text. The language model used was GPT-3, developed by OpenAI, and uses natural language processing and deep learning to produce text based on provided instructions. The model was given the following instruction: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." Below is the second paragraph of the machine-generated article. (GPT-3, 2020).

> *The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.*
>
> (GPT-3, 2020)

Besides GPT-3, there are other pre-trained NLP models that can be used. BERT, developed by Google, can also be used in text generation. Both GPT-3 and BERT are pre-trained with billions of parameters and can be fine-tuned for more specific tasks. (Devlin et al, 2019 & Brown, et al, 2020).

Text-generating NLP models can also be trained from scratch as long as there is a sufficient amount of data to train the model against. Below is a simple flowchart on the steps involved.

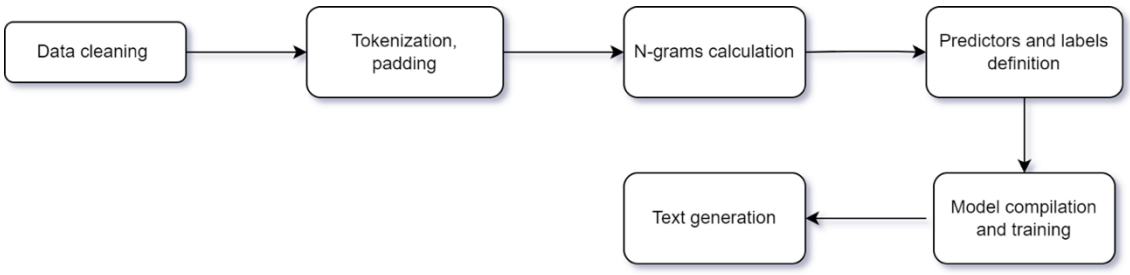

Figure 3. Flowchart of steps to create a NLP model (Korab, 2022)

This thesis does not dive very deeply into the technical aspects of training neural networks or explain in great detail how machine learning algorithms work. These models require very large quantities of data to provide accurate results, as opposed to humans, who "do not require large supervised datasets to learn most language tasks – a brief directive in natural language or at most a tiny number of demonstrations is often sufficient to enable a human to perform a new task to at least a reasonable degree of competence" (Brown, et al, 2020). Also, that the language is good does not mean the facts are correct.

Of all the forms of automation in news production, machine learning is still the least used, based on its complexity and need for resources, both computational and material; a vast archive of articles is needed to train machine learning models against. Especially automated processes that are deployed in practice tend to be made by templates rather than machine learning algorithms. (Kanerva, et al., 2019, p. 1).

3.1.3 Virtual Assistants and Article Curation

AI-powered tools can help journalists recognize correlation and causal links, or interesting outliers, in data by automatically flagging them for verification (Marconi, p. 38).

Many of the tools employed by journalists and integrated in their outlets' computer systems are a form of automation: for instance, widgets integrated into the outlet's content management system (CMS), that can provide automatically generated maps of where the article's events take place, or a chart of how MP's or parties voted on a specific issue.

Another type of automation that helps a journalists' workflow is a browser extension, that can be configured to alert the journalist when a certain web page (or part of it) has changed, thus freeing up the journalists' time. (Salmela, 2019, p. 33)

AI algorithms are increasingly also being used to curate the web sites of online publications, to automatically decide which articles to highlight, or when to put an article behind a paywall (Future Today Institute, 2022 (a)).

Other uses are for instance automatic keyword extraction, where an article is automatically analyzed, and a set of keywords are provided. They can then be used for classification or search functions. Yle has taken into use Annif, an open-source software developed by the National Library of Finland, that uses a combination of NLP and ML methods. Yle has integrated it into its publishing systems and has trained it with in-house articles in Finnish and Swedish, and updates the vocabulary on a weekly basis. (Kauranen, 2021).

A step more advanced than keyword extraction is automatic summarization of articles, where an article is made shorter or customized for a different format, while preserving the main message. The summaries can be used for different purposes, such as snippets of an article on the main page or in search engine results, or to change the format into listicles or shorter, mobile-friendly posts or even notifications (Marconi, 2019, p. 42). The summarization can be performed with NLP algorithms, by extracting sentences without modification (*extraction based*) or by generating new sentences based on the full text (*abstraction based*) (Lehto & Sjödin, 2019).

3.2 Real-World Examples

Artificial intelligence, through machine learning, has so far not lived up to its potential in real applications. “So far, robot-made articles everywhere have been based on templates that the program fills in. They do not have machine learning”, says Carl-Gustav Lindén, Assistant Professor at Helsinki University (according to Vehkoo, 2017). Hence, most of the examples below are template-based, and primarily focused on fields of journalism that are data-heavy, ie. have by nature an abundance of quantifiable data points, as opposed to for instance human interest stories that rely on interviews or the journalists’ experience of a location or event.

There are numerous examples of automation in different fields of journalism. Automation requires large amounts of data in a structured form, hence fields where there is an abundance of data – sports, business reports, elections, weather – are the first to be automated.

There are many companies that develop and sell customized platforms for news outlets.

3.2.1 Sports Journalism

An early example is StatsMonkey, which started as a research project from Northwestern University, and automatically wrote recaps of baseball games. “Baseball served as an ideal starting point due to the wealth of available data, statistics, and predictive models that are able to, for example, continuously recalculate a team’s chance of winning as a game progresses.” (Graefe, 2016, p. 19).

MittMedia, a Swedish media company, publishes 3 000 automated texts per month on soccer games – from all levels of matches. Employees call team leaders and referees by phone to collect match events and quotes, but have also introduced a chat bot to collect usable quotes automatically. (Lindén, et al, 2019, p. 15).

In 2016, Yle developed a template-based robot called Voitto that produced sports news. It first started by writing ten articles about ice hockey games in the NHL (Hallamaa, 2016), and then moved on to to cover other kinds of sports, such as floorball and football. (Ryynänen, 2021, interview). Voitto’s template model is presented before, and more details are provided later in this thesis.

3.2.2 Financial Journalism

Using templates to automate some of its financial news stories, the Associated Press has gone from having human journalists cover 300 companies to having machines cover 4,400 companies (Marconi, 2019). Automation in financial news is seen as a must in order to compete: “Thomson Reuters and Bloomberg extract key figures from press releases and insert them into pre-written templates to automatically create news alerts for their clients. In this business, automation is not about freeing up time. It is a necessity.” (Graefe, 2016, p. 20).

3.2.3 Political Journalism

Automation has already changed the way elections are reported on. For instance, using automation tools, the Washington Post was able to cover the 2016 US elections in all states, including 435 house races, 34 senate seats, and 12 gubernatorial races. In previous elections the articles were written by human journalists, who could only cover 15 percent of the different races. (Diakopoulos, 2019, pp. 109–110)

In Finland, the team at research project Immersive Automation created a news bot that reported on the 2017 municipal elections in three languages: Finnish, Swedish and English. Compared to a human journalist, the bot was incredibly efficient. “Valtteri produced more than two million news stories across all three languages”, each of which would have taken a journalist one hour to write. (Lindén, et al, 2019, p. 22). The calculated difference in output is staggering, although the usability of two million articles on elections in a country of five million can be questioned.

3.2.4 Weather Reports

Weather reports are arguably the first field of journalism in which automation was used, over 50 years ago. An early study describes a piece of software that takes the outputs of weather forecasting models, prioritizes them and uses pre-written phrases to generate “worded weather forecasts”, which in effect is a good example of template-based automation. (Graefe, 2016, p. 20)

3.2.5 Other Types of Journalism

Automation is used by newsrooms to tackle other types of issues as well, often as custom solutions. For instance, the Los Angeles Times have automated homicide and earthquake reporting (Graefe, 2016, p. 20). Similarly, after its use on elections was over, Immersive Automations’ Valtteri bot was turned to focus on crime using publicly available statistics (Lindén, et al, 2019, p. 22).

3.3 Impacts of Automation

> *What constitutes "newsworthy" changes when it's cheap and easy to cover basically everything.*
>
> (Diakopoulos, 2019, p. 110)

When a suitable automation is developed, the cost per article goes down for each article generated. In such a situation, it is “worth it” to generate articles even on niche topics and issues that are not interesting to the masses. Graefe (2016, p. 22) argues that there are two obvious economic benefits to automating the data collection, writing and publication of news stories: increasing the speed and scale of news coverage.

The risk with automating many similar stories, especially with templates, is that the articles are very similar from day to day, and more so if the data sources are open and in use by several news organizations. The outlets that have access to unique databases may very well get a competitive edge, and investing in exclusive data sources will most likely increase in the future. (Diakopoulos, 2019, p. 117).

The obvious impacts are divided into optimistic and pessimistic views, two sides of the same coin if you will. Optimists see that automation will do the boring work, allowing the journalists to focus on tasks that the machine cannot: to conduct interviews, edit and enrich machine-generated drafts, and to put more thought into the perspectives and narrative. Pessimists view automated journalism as a genuine threat to their lifestyle and livelihood. (Latar, 2019, pp. 29–30).

> *Algorithms do not get tired or distracted, and—assuming that they are programmed correctly and the underlying data are accurate—they do not make simple mistakes like misspellings, calculation errors, or overlooking facts.*
>
> (Graefe, 2016, p. 23)

Robot journalists always operate according to a predefined ruleset, and as such are not susceptible to typos or other errors – but there might be human errors in designing the templates algorithms fill, for instance a field might be mapped wrong so the right data is in the wrong place.

The speed at which news automation becomes more widespread in newsrooms is probably overestimated. Professor Andreas Graefe commented in 2018 that “Five years ago, there were many bold predictions about how automated journalism will develop. From claims that 90% of news will be automated to Pulitzer prizes for automated content. In reality, not much has changed. Progress is steady but slow.” (According to Lindén, et al, 2019, p. 44).

One reason news automation is slow to evolve could be its low priority among news leaders (246 were asked to fill in the survey). In Reuters Institute’s yearly Journalism, Media, and Technology Trends and Predictions for 2022, robo-journalism rated least

important use of AI, behind recommendations and commercial uses. The tools that help journalists write news, like virtual assistants described above, was deemed more important. News automation seems to not be mature enough for wider adoption, "but is where many of the most future-focused publishers are spending their time." (Newman, 2022).

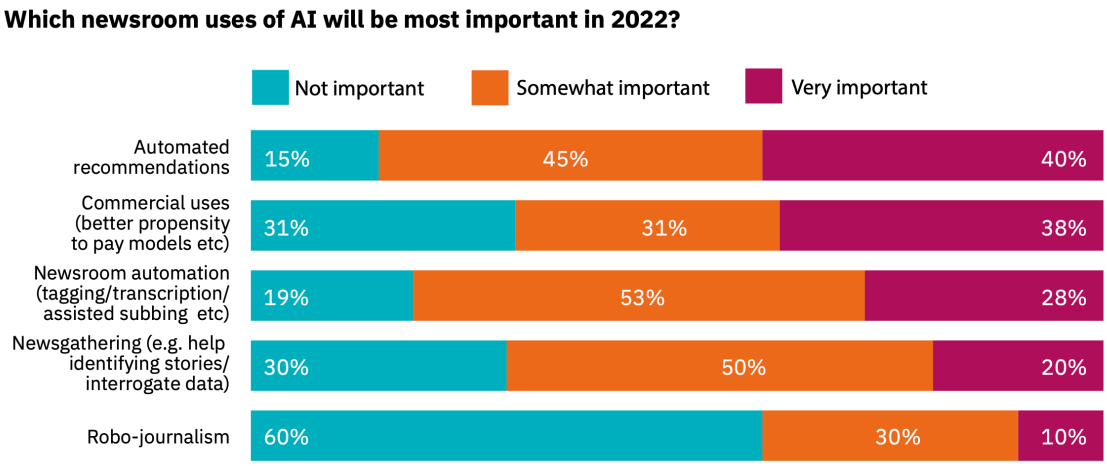

Figure 4. Bar chart of importance of AI use by category in newsrooms (Newman, 2022)

3.4 Ethics Considerations

There are many ethical questions to be considered when using algorithms and artificial intelligence to produce news.

The Council for Mass Media in Finland (CMM) is a self-regulatory organ of the Finnish publishers and journalists, which publishes the Journalist's Guidelines, a collection of the ethical rules of journalism in Finland that most outlets adhere to. (Ala-Fossi, et al, 2021, p. 162).

In 2019, the Council for Mass Media issued a statement on how the use of automation and personalization of news should be marked. At the same time, it also defines that the use of algorithmic tools is indeed journalistic work, and that media outlets should act responsibly and transparently while using algorithms. (Council for Mass Media in Finland, 2019).

When news organizations are planning to take automated content generators into use, they also need to consider some legal factors, such as copyright infringement or libel as a consequence of automation (Lindén, et al, 2019). Even if the error is made by a machine, the decision to publish is still made by humans. It is not possible to sue an algorithm, so humans are still the ones responsible (Broussard & Lewis, 2019).

3.4.1 Black Boxes

Newsrooms often lack the necessary resources and skills to be able to develop automation tools themselves, forcing them to collaborate with companies that develop software. (Graefe, 2016, p. 20). This has the potential of putting the transparency at jeopardy.

The Council for Mass Media stipulates that the power of journalistic decision-making should not be transferred to companies making the algorithms, and that developers of these automatizations also adhere to the guidelines for journalists. The responsibility for ensuring that the decision-making power is not relinquished lies with the editorial office and the chief editor, who must understand the effects of algorithms. The Council's statement is provided below in its entirety:

> *The Council states that the use of news automation and targeted content always constitutes a journalistic choice. It includes choices about what to publish, with what emphasis and to whom. This is at the core of journalistic decision-making power.*
>
> *The guidelines for journalists stipulate that the decisions concerning the content of media should be based on journalistic principles and the power to make such decisions should not be surrendered to any party outside the editorial office. Consequently, such decision-making power should also not be transferred to makers of algorithms outside the editorial office. The editorial staff – ultimately the chief editor – bears the responsibility for the effects of algorithms on journalistic content.*
>
> *The Council states that media outlets should have sufficient understanding of the effect of algorithmic tools on content. For example, if a media outlet purchases a tool developed externally, the outlet must examine and approve*

> *of its central operational principles and be able to react, should problems arise.*
>
> *The Council reminds that the guidelines for journalists apply to all journalistic work. Media outlets must therefore ensure that their digital service developers also adhere to the guidelines, when they independently make decisions that influence journalistic content.*
>
> (Council for Mass Media, 2019)

When it comes to news produced by an algorithm or a machine learning model, the grounds on which they operate are not always easily understandable. This is called explainability, where even if some machine learning algorithms can be very accurate in producing results, but very opaque in terms of how they make decisions. "The notion black-box AI refers to such scenarios, where it is not possible to trace back to the reason for certain decisions. Explainability is a property of those AI systems that instead can provide a form of explanation for their actions." (High-Level Expert Group on Artificial Intelligence, 2019).

In other words, template-based automations are code that can be traced back to see why and how the template made choices. machine learning models are trained on vast amounts of data, and the underlying rules cannot easily be made visible and the results traced back to how they were reached.

3.4.2 Who Gets the Byline?

> *The byline, that is the line that states who wrote or collaborated on the article, is a standard practice to give credit where credit is due. In the case of automated journalism, the practices were heavily varied between different online publications. For example the Associated Press notes how the article was generated on the bottom of every automated news story (Marconi, 2019, p. 97).*

In Finland, the Council for Mass Media expressed its view that the public has a right to know about news automation, and obligates its member outlets to "disclose to the public if journalistic content published by them has, to an essential extent, been generated and

published automatically", and recommends that media outlets disclose the type of automation and the source of the information. (Council for Mass Media, 2019).

4 EMPIRICAL FINDINGS

The following findings are based on interviews conducted with experts from Yle and STT.

4.1 Template-based Automations

At Yle, the Voitto robot is still in active use, although it has evolved since being introduced in late 2016. It's current use is described in its own chapter below.

At STT, there have been several projects related to automation, but there are no systems actually producing news articles currently in use, mainly because of a lack of resources. "If there are templates, there always needs to be someone who maintains and produces and makes them." (Interview with Salmela, 2021). Yle has even provided STT with a version of its own Voitto robot, but it is not in use either (Interview with Ryynänen, 2021).

4.1.1 Case: Voitto

The Voitto robot, of which the open-sourced version's templating functionality has been described in a previous chapter, has been developed further and an improved version of it is still in use at Yle. Voitto's first assignment in December of 2016 was to write match reports on ten NHL ice hockey games overnight, which it successfully did. Since then, it has written stories on different sports – more ice hockey, floorball, football and others – but not NHL, as those were games human journalists wanted to cover (Interview with Ryynänen, 2021). It's quite interesting that journalists' preferences affect what kind of assignments robots get – and that humans get to cover the big leagues, while less prestigious games' reporting is found suitable to automate.

After a successful start in sports, Voitto started preparing for the Finnish municipal elections in the spring of 2017. That was also a success, so since then Voitto has been writing about politics quite a lot. "Nowadays, Voitto follows politicians. We are gathering information about Members of Parliament. And we make once a month a newsletter that tells what Voitto has done in the past month." (Interview with Ryynänen, 2021).

The technology that Voitto is built upon is built upon is a lot of short template texts, which are then filled with relevant data points based on available data for the given topic. The idea is that almost anything that has associated, machine-readable data can be automated. That is why sports is an easy choice, there is a lot of time-stamped match data available. The same applies for politics, and that can be even more interesting to follow.

> *It's not very dangerous to automate sports news because if you make a mistake, it's not so bad, but on a political area, if you make a mistake, it could be a scandal. So therefore, it's way more interesting to work on a political area, but it's also more dangerous.*
>
> (Interview with Ryynänen, 2021)

Voitto is able to report on elections, how a Member of Parliament has voted, or what they have said during assembly or meeting attendance. “We are trying to gather all kinds of statistics about Parliament Members’ doings and sayings, but for now we haven’t gone beyond that.” (Interview with Ryynänen, 2021).

In addition to writing about politics, Voitto has learned to draw. Whenever there is a vote in the Finnish Parliament, Voitto draws an image of how the votes were distributed among the parties and MP’s, so journalists can use a ready-made image in their news article. (Interview with Ryynänen, 2021).

The problem with template-based solutions is that the public will find fully automated news repetitive. This is also one of the principal reasons why automated news needs to be edited by a human journalist before publication. The robot can create the basis for the article, and a human journalist can enrich it by adding descriptive text, providing background and context, or for instance by conduct interview related to the topic. This way, articles can be quickly published by automation, only to be enriched with more content by a human journalist. To begin with, Voitto’s news were completely automated, because the point of automation is to free up human resources. But Voitto has evolved, and nowadays only writes a small number of articles and a monthly newsletter, but they too are edited and enriched by a human journalist before publication. (Interview with Ryynänen, 2021).

4.1.2 Valtteri

The Valtteri bot detailed earlier was in development around the same time as Voitto, but did not get traction at Yle. “It was a different project. We talked about maybe cooperating, but because of political issues between a government media versus commercial media, we didn't find any any clear land to work together, so to say, so therefore they are separate projects.” (Interview with Ryynänen, 2021).

4.2 Machine Learning

Ryynänen notes that “Voitto is nothing about machine learning or artificial intelligence” and that only the recommendation engine has a machine learning component. “We haven’t been trying to make synthetic news, so to say, that [a] machine wouldn’t have any kind of templates, but that [an] algorithm would somehow learn to write human language. […] But those are kind of top edge technologies anyway in this planet so therefore we haven’t been trying those in Finnish yet, but maybe one day.” (Interview with Ryynänen, 2021).

4.2.1 Case: Scoopmatic

STT in turn has tried developing a cutting-edge machine learning model for sports news in project Scoopmatic, funded by the Google Digital News Initiative and done in collaboration with researchers from the University of Turku. “It first learned to write Finnish in the STT style using our own digital archives, and then we started giving it parameters to write about. Our aim was to create a model that could write news on ice hockey. It would be given the match data and based on that and what it knew from our previous hockey news, it would write the text. And it could do it. Its Finnish was decent.”
The problem was that it could suddenly invent events that did not happen, such as write about a team that was not on the ice in that match, or use language out of context.

> *It was really interesting to see that with relatively little material to teach it with it learnt such a small language as Finnish, and that the output is much more interesting for the reader than a recurring template.*
>
> *But when we talk about journalism where facts matter, we didn’t reach the point where it would be factual enough to publish directly. Or that it would*

have been so factual that a journalist would not have had to edit each short text and clear it for publication.

If we have that kind of a solution, it won't save us any of the resources and time that we would want it to save. But this was an interesting experiment.

(Interview with Salmela, 2021)

The Scoopmatic project dataset comprised of 3,454 Finnish leagues' ice hockey games with statistics and at least one corresponding news article, from 1994–2018. (Kanerva, et al, 2019, p. 2). The problem with machine learning is that the need for learning material is immense, more than was available in STT's archives. "You need like millions of items. […] They tried to make sports news and there is simply not enough sports news in Finnish to learn that kind of algorithm correctly." (Interview with Ryynänen, 2021).

The problem with these models is, if you think about text generation, that if they hallucinate, then STT can't use them. Trustworthiness is our principal value. If our news can't be trusted, we don't have anything, and we don't publish any best guesses.

(Interview with Salmela, 2021)

The Scoopmatic project has ended, but the developed model is publicly available. (Interview with Salmela, 2021).

4.2.2 Summarization

Yle has been trying summarization, but the quality has not been good enough to publish condensed versions of articles without review. They can be classified as helper tools for journalists. (Interview with Ryynänen, 2021). However, the keyword extraction using Annif is based on machine learning and trained with Yle articles instead of The National Library's material. (Kauranen, 2021). STT has so far not used versioning or summarization of its articles. (Interview with Salmela, 2021).

4.3 Resources

With regard to resources, Yle and STT play in different leagues. At Yle, transparency is requiring them to develop their own tools instead of purchasing third-party software, as there is no transparency into how software developed by third parties works.

STT lacks the resources for any in-house development of automation. To counteract this, they take part in cooperation projects, such as the Google DNI -funded Scoopmatic described above, which had University of Turku researchers create a machine learning model based on STT's news archive (Paikkala, 2020), or the EU's Horizon 2020 -funded EMBEDDIA project, during which a handful of tools were developed and released (Pollak, et al, 2021), and with Finnish media when there are mutual interests and benefits. (Interview with Salmela, 2021).

Templates always require someone who programs, updates and understands them, and if that person is not a part of the newsroom, it is hard to train a journalist to do it on the side. "The primary goal of automation is to free up journalists' time to do something more meaningful. If all the time it takes going over sports results or finance reports would be freed up for more in-depth reporting, then that would be better service for our customers and the audience who the news are for in the end." Also, through automation STT hopes to offer content that they currently cannot do, such as automated sports news from regional leagues. (Interview with Salmela, 2021).

In an effort to save journalists' time, STT has developed a tool called Pikkulintu ("small bird" in Finnish), that is a browser extension that can be asked to monitor web pages for changes or keywords, and then it alerts the journalist. That way they don't have to spend time manually checking for updates, but can react quickly when needed. (Interview with Salmela, 2021).

4.4 No More Black Boxes

> *We actually build all the tools by ourselves, or pretty much.*
>
> (Interview with Ryynänen, 2021)

To tackle the issue of black boxes, Yle builds its own tools itself. That way they do not have to guess how a software bought from a private company works.

STT's Salmela (2022, interview) raises the same issue: very few news organizations in Finland have the needed resources, even if there is a demand for certain types of automated software.

The issue that Salmela raises about the difference in troubleshooting a template-based techniques and machine learning techniques is also more generally recognized in AI.

Yle, on the other hand, is taxpayer-funded and as such does not have the same kinds of issues with resources. Jarkko Ryynänen (interview, 2021) says that as a public broadcaster, Yle must know how the automation tools producing the news work, so as to avoid any kind of bias by the robot journalist. That is why Yle develops its own tools in-house instead of buying or licensing ready-made tools. It has previously also used commercial products but is moving away from that to be in control of its own tools.

Unlike templates, one cannot see inside a machine learning model to see on what basis a certain decision is made and why, essentially making the process a black box. This can, at worst, lead to completely unforeseen results in an article, such as radical events that have not actually taken place.

> *If a template-based machine makes a mistake in an article, then you just go in and see inside the template and locate the error in the code. But if Scoopmatic makes a text, and this is a real example, where an ice hockey game starts, but at the end everyone is dead on the ice, it's hard to ask the machine why it made these choices. My assumption is, that if you look at our whole archive, the news are often about conflicts and the language of conflict and sports is quite similar.*
>
> (Interview with Salmela, 2021)

Salmela calls these kinds of machine learning issues as the AI "hallucinating", to which a possible solution could be to use templates in conjunction with machine learning algorithms. (Interview with Salmela, 2021).

4.4.1 Robot Bylines

Both Yle and STT adhere to the Council for Mass Media's guidelines for journalists, although there are slight differences in their approach. Salmela says that "if it's an NLP model that writes the first version of a sports or finance story, then there has to be a mention that it's automated text, and a short explanation of what it's based on, but in our case the credits would still go to STT".

At Yle, the byline is shared in case there is a human journalist involved: "Voitto gets [its] own byline and those human beings involved get their own so everyone is mentioned. So Voitto is like one of the colleagues of journalists, so they are in the same level." If the article includes images or widgets that are compiled by Voitto, it is mentioned inline, but not at the end of the article. (Interview with Ryynänen, 2021).

4.5 Responsibilities

Using automation in the newsroom raises questions about responsibilities and how they are divided. The automated machine could be connected to a bogus data source, and if news based on this are published automatically or with too few checks, then it could go against the guidelines for journalists, or even constitute an offence, such as libel. "Machines don't notice these things, so a human must to some extent monitor what the machine is doing." (Interview with Salmela, 2021).

Systems errors do happen, or errors in the code, which might lead to a news article not being entirely true. In that case, it's most important to correct these mistakes and openly tell the audience what happened. (Interview with Salmela, 2021).

At Yle, the responsibilities are one of the main factors why they use templates. "In the end, there's always a human being who is responsible for those articles. And usually those human beings want to know how this machine works." (Interview with Ryynänen, 2021). Whether to publish or not, or whether to automatically publish automated stories

or not, is still a human decision. From that standpoint it doesn't matter if it's a human journalist or an automated one who has written the article – errors are handled in accordance with a normal journalistic process, where factual errors are corrected, no matter who made the error. (Interview with Ryynänen, 2021). The responsibility lies with the publisher.

> *So when you're using templates, you can say that if we put in this kind of data, we will get this kind of output. But then machine learning, it's more complicated. You can say that if we put this kind of data in, we get pretty much something like this, but we can't be sure. And usually that's not the answer that [the] human being responsible wants to hear.*
>
> (Interview with Ryynänen, 2021)

"For some reason we do not trust machines as much as we trust human beings. And maybe there's a good reason for that. Usually people are more afraid of [a] machine doing bad things than human beings doing bad things. [Machines] are so fast that they can do so many bad things in such a short time period, so there is a point in that way. (Interview with Ryynänen, 2021).

4.6 Reception by Journalists

A natural urge of humans is to think of as rivals anything that does our work faster or more efficiently. That is the case with AI as well. STT's Salmela (2021, interview) thinks the resistance is not so much against the new tools, but against having no time to properly learn them.

> *I've often been told that us journalists have many doubts and are very hesitant to change. I don't agree. I'm not sure if it's because I work in a news agency, but we have very many workflows where this [automation] suits very well, and people generally are quite excited about [for instance] getting financial data directly into the content management system. They understand that there's still so much work to be done that if the financial data they need comes*

automatically, they can't go to lunch for the rest of the day, but instead get five minutes to do other stuff.

(Interview with Salmela, 2021)

That is to say that what is perceived as a negative reception towards automation might in fact be misdirected, and the real cause for skepticism is the lack of support in taking the new tools into use. "There is hesitancy, but mostly it's about people understanding the realities of their own work and workflows. So my experience is that if people think they don't get enough orientation, if they need to start using a new tool but don't know how, then it feels like it makes the work harder." (Interview with Salmela, 2021)

In the interview, Ryynänen tells the birth story of Yle's Voitto robot almost like an anecdote. "The reactions are pretty interesting. When we started this thing in 2016 and Voitto wrote ten articles of NHL games. And we were so proud when we went to the sports department to show what we have done. And all the journalists were afraid of what kind of monster this Voitto is and the first or second question was that 'will this robot take my job?' So that was their major issue. They weren't impressed at first that what this kind of automatic thing could do."

The gut reaction by journalists was first fear for their own employment, and the development team was suspected of being in on a scheme to reduce the number of people in the office. The first articles were credited to Voitto (which has a dual meaning in Finnish, the literal translation is "victory", but it is also a male first name) and an image of a robot. For some reason, the development team then decided to make a plush doll of Voitto, a sort of 1-meter-tall marketing gimmick.

People started to think that it's actually the doll that makes those stories, and since that doll was kind of cute, so they weren't afraid of that robot anymore, that's [a] very strange thing, and it was a [sic] lucky that we decided to make that doll. I can't remember why we [made it], maybe it was because we have to have something to show because it's kind of abstract to show some kind of lines of code. But it was [a] lucky accident. So people, instead of being afraid of that robot, they wanted to take selfies with that doll.

(Interview with Ryynänen, 2021)

After that doll, the attitudes at Yle changed. “People weren’t afraid anymore and they started to think that ‘I have so much dull work that that thing could do, and I could do something more interesting’” (Interview with Ryynänen, 2021).

Salmela sees that the news industry’s understanding of automation has increased in the past five years. Journalists are no longer as worried that robots will come and take their jobs. “They are starting to understand that if robots take their jobs, the journalism that is left is not going to be terribly good”. (Interview with Salmela, 2021).

4.7 Reception by the Audience

According to Ryynänen, the Voitto-generated articles are met with a quite the binary reaction, some people like them very much and other hate them. “They actually said that they won't read anything from Yle anymore because there is a robot doing human’s job.” On the other hand, some people love traditional teletext pages because of the simplified, no-nonsense format, and Voitto sports reports “are kind of teletext pages but maybe version 2.0. There is more stuff in that [article], but basically the same thing that it's [a] simplified story of a match, how it went, [and] what was the end result.”

As news agencies lack their own publication channels and the articles are published by the agency’s customers, STT has a more limited insight into how articles written by them are received. “We don’t have direct visibility into article comments where people might leave their first gut reaction.” (Interview with Salmela, 2021). For the time being, STT has no automated news article systems in production.

4.8 Future Outlook

How does the future of automation look like in the next five to ten years? Yle’s Ryynänen has high hopes for machine learning models. “We will have that synthetic text producing in Finnish in five years, I'm quite sure of that.” A machine might even win the Pulitzer Prize for the first time. (Interview with Ryynänen, 2021).

Salmela believes that more and more automation applications are taken into use in newsrooms, as long as they are cost effective, reliable and user friendly. “If they [the

automation tools] stick to the facts and don’t hallucinate, then they will be taken into use. […] But if the articles need to be proofread by a journalist before publication, then a human might as well write the article themselves.” (Interview with Salmela, 2021).

5 CONCLUSION

> *Newsrooms see content automation as being largely complementary to journalists' work. Yes, there are instances in content production where there is complete automation, and if you squint, you might even say there is artificial intelligence operating in narrow targeted areas. But the state of the art is still far from autonomously operating in the unbounded environment of the world and from doing the contextualized interpretation and nuanced communication required of journalists.*
>
> (Diakopoulos, 2019, p. 97)

So what can be said about the stage of automation in newsrooms? There has, for the past decade, been a steady increase in the usage of automation in the newsroom, starting from finance, sports and weather, fueled by the increase of quantifiable data through datafication, the media's shift to online publication, and the development of templating and automation tools, and in parts through advancements in the field of neural networks, machine learning, and natural language processing.

While this progress has been seemingly fast, this intelligent automation simply has not been able to live up to the wildest predictions: autonomous robot journalists, Pulitzers, and 90% of news being written by machines. There have been many trials and errors.

The examples from the organizations this thesis has most focused on, Yle and STT, have had very different kinds of projects that are almost hard to compare, but in a way neither of them have yet managed to create a system that is completely autonomous, despite the initial idea that a machine's work must be so reliable, trustworthy and accurate – and fulfil the outlet's criteria on language – that it doesn't require human intervention, or at least not too much of it time-wise.

Yle's Voitto robot started out as a simple template-based system, writing sports and election news that could be published automatically. While it performed adequately, the fact that is that template-based systems tend to become repetitive – there are only so many ways a story that needs to include the same key factors can be written. Voitto has since evolved, and now works in tandem with a human journalist. Voitto can provide

the base, to which a human journalist may build upon by doing something only a human – at least now – can do: provide background, analysis, and interviews. The sports articles Voitto writes are built on statistics, time-stamped events from the game's logs. There is no deeper understanding or emotion that can be conveyed to the reader.

STT's Scoopmatic project, the pilot into producing news using machine learning algorithms, thus achieving richer and more diverse language is commendable. In this case, it seems the material with which to train the model was simply not sufficient, even though it encompassed all of STT's archived articles on ice hockey matches in Finnish leagues. While the resulting articles were interesting and at times hilarious, they were simply not ready for publication, as the events described did not actually take place. To be published, they too would be in dire need of a human journalist to correct the errors. Insufficiently trained, the machine makes up its own events. This tendency to hallucinate could possibly be counteracted by using a hybrid model, where both templates and machine learning would be used for a single article.

Voitto has now found a way to work as a tool for journalists, but Scoopmatic is for now not being used or further developed at STT. Both projects have served an important role, though: in the early days of automated news, journalists were afraid of having to compete or even of losing their jobs to these new machines, but now most human journalists accept that they are more like another tool in the toolbox: automating the tedious part of newsgathering or article generation, and giving the journalist more time to focus on what they do best: write compelling, thought-out stories, conduct interviews and provide the angle and emotion in a story. After all, it is still the human who is responsible for what is published, so it must be the humans who tell robots what to do.

In this way, the thinking and attitudes towards automated journalism has progressed quite a lot. The adoption of these systems is not only to automatically create more news, but to increase the quality and speed with which humans can write news. That, in effect, requires accepting to shift the objective from machine autonomy to machine collaboration, but if the end result of better news faster is reached, does it really matter?

We can safely say that complete automation of news is still a ways off. Template-based systems are being used in newsrooms, but often as a basis for a journalist to expand

upon. Machine learning is mostly deemed too unreliable to be used for writing articles, but has found its use as a tool for suggesting keywords, correcting spelling or looking up facts as the human journalist writes an article. So maybe the future is not about robot journalists replacing human journalists, but instead about robots working alongside humans, providing more advanced tools.

Even though a deeper delve into the topic is outside the scope of this thesis, it warrants at least a passing mention in the conclusion: As digital data is increasingly used for news production, it is also used to track and analyze the behavior of readers of online publications, as well as on social media. When readers are sufficiently profiled, it opens up the possibility for not only recommended articles on a personal level, also automatically tailored news for each reader – such as different articles on an ice hockey match, based on which team the reader supports, or on a range of other topics – which could have far-reaching effects on societies as a whole, possibly driving polarization and news bubbles, when the tone of news can be customized per reader using the same tools used for automating the news as a whole. This risks proving true, even for journalism, Zuboff's third law: "In the absence of countervailing restrictions and sanctions, every digital application that can be used for surveillance and control will be used for surveillance and control, irrespective of its originating intention."

Robots do not – at least not yet – consume news the same way humans do, so they cannot be entrusted with the autonomous publication simply because they lack the human condition of emotion, without which there cannot be a complete grasp on context. No matter how thoroughly the model has been trained, even eloquent writing is just an emulation and amalgamation of human journalists' collective effort.

Progress for the actual intelligent automation of news has been slower than anticipated. But as new machine learning algorithms are created and old ones improved, the situation may quickly change. OpenAI's GPT-4 is expected to come out shortly after this thesis is published, and offer vastly increased performance. News outlets' interest in pursuing news automation and allocating resources for their development is a limiting factor, as according to surveys (see p. 33) they are prioritizing other uses for AI in the newsroom.

It remains to be seen how the intelligent automation of news is taken into use in newsrooms both in Finland and abroad. As of now, the technology is not mature enough for articles to be written completely autonomously, and the newsrooms are responsible enough to at a minimum double-check what the machine has written, but in many cases collaborate with the machine, combining the skills and abilities of human and robot journalists to produce more and better journalism, and both are mentioned in the byline. This is ultimately to the benefit of the audience, the readers and consumers of the news.

REFERENCES

Ala-Fossi, Marko, Grönvall, John, Karppinen, Kari, & Nieminen, Hannu. "Finland: Sustaining professional norms with fewer journalists and declining resources". In J. Trappel, & T. Tomaz (Eds.), *The Media for Democracy Monitor 2021: How leading news media survive digital transformation (Vol. 1)* (pp. 153-196). Nordicom, University of Gothenburg, 2021. http://norden.diva-portal.org/smash/get/diva2:1557246/FULLTEXT01.pdf. Retrieved 12.6.2022.

Beckett, Charlie. *New powers, new responsibilities: A global survey of journalism and artificial intelligence*. The London School of Economics and Political Science, 2019. https://drive.google.com/file/d/1utmAMCmd4rfJHrUfLLfSJ-clpFTjyef1/view. Retrieved 12.6.2022.

Brinkmann, Svend. *Qualitative Interviewing*. Oxford University Press, 2013.

Broussard, Meredith & Lewis, Seth. *Will AI save journalism – or kill it?* Knowledge@Wharton 9.4.2019. https://knowledge.wharton.upenn.edu/article/ai-in-journalism/. Retrieved 12.6.2022.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh; Aditya; Ziegler, Daniel M.; Wu, Jefrrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario. *Language Models are Few-Shot Learners.* OpenAI, 2020. https://arxiv.org/abs/2005.14165. Retrieved 12.6.2022.

Cognizant. *Glossary of Terms: Intelligent Automation*. Cognizant, 2022. https://www.cognizant.com/us/en/glossary/intelligent-automation. Retrieved 12.6.2022.

Collins English Dictionary. *The dictionary entry for "neural network"*. HarperCollins. https://www.collinsdictionary.com/dictionary/english/neural-network. Retrieved 12.6.2022.

Council for Mass Media in Finland. *Statement on marking news automation and personalization.* 2019. https://www.jsn.fi/en/lausumat/statement-on-marking-news-automation-and-personalization/. Retrieved 12.6.2022.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton & Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Google, 2019. https://arxiv.org/abs/1810.04805. Retrieved 12.6.2022.

Diakopoulos, Nicholas. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures". *Digital Journalism*, 3(3), 398–415, 2015.

Diakopoulos, Nicholas. *Computational Journalism and the Emergence of News Platforms.* 2016. http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Computational-Journalism-and-the-Emergence-of-News-Platforms.pdf. Retrieved 12.6.2022.

Diakopoulos, Nicholas. *Automating the News: How Algorithms are Rewriting the Media*. Harvard University Press, 2019.

The European Commission. *European legislation on open data*. European Commission, 2022. https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data. Retrieved 12.6.2022.

The European Commission's High-Level Expert Group on Artificial Intelligence. *A Definition of AI: Main Capabilities and Scientific Disciplines*. European Commission, 2018. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf. Retrieved 12.6.2022.

Federman, Mark. *What is the Meaning of the Medium is the Message?* 2004. http://individual.utoronto.ca/markfederman/MeaningTheMediumistheMessage.pdf. Retrieved 12.6.2022.

Future Today Institute. *Tech Trends Report, Vol. 05: News & Information*. Future Today Institute, 2022 (a). https://futuretodayinstitute.com/mu_uploads/2022/03/FTI_Tech_Trends_2022_Book05.pdf. Retrieved 12.6.2022.

Future Today Institute. *The Global Survey on Journalism's Futures*. Future Today Institute, 2022 (b). https://futuretodayinstitute.com/mu_uploads/2022/03/GlobalSurveyJournalismFuture_031422-1.pdf. Retrieved 12.6.2022.

Galily, Yair. "Evolution, Revolution or a Real Game Changer? Artificial Intelligence and Sports Journalism". *Robot Journalism: Can Human Journalism Survive?* World Scientific, 2018.

GPT-3. *A robot wrote this entire article. Are you scared yet, human?* The Guardian, 8.9.2020. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3. Retrieved 12.6.2022.

Graefe, Andreas. *Guide to Automated Journalism*. Tow Center for Digital Journalism Publications. Columbia University, 2016. https://academiccommons.columbia.edu/doi/10.7916/D80G3XDJ. Retrieved 12.6.2022.

Hallamaa, Teemu. *Voitto-robotti takoi NHL-uutisia – seuraavaksi kuntavaalien tulokset?* Yle, 28.12.2016. https://yle.fi/uutiset/3-9375528. Retrieved 12.6.2022.

Hamilton, James T. & Turner, Fred. *Accountability through algorithm: Developing the field of computational journalism*. Stanford University, 2009. https://web.stanford.edu/~fturner/Hamilton%20Turner%20Acc%20by%20Alg%20Final.pdf. Retrieved 12.6.2022.

Haugeland, John. *Artificial Intelligence: The Very Idea*. M.I.T. Press; Bradford Books, 1985.

Humby, Clive. *Data is the New Oil*. ANA Marketing Maestros, 2006. Edited by Michael Palmer. https://ana.blogs.com/maestros/2006/11/data_is_the_new.html. Retrieved 12.6.2022.

Hyppönen, Henkka. *Luomiskertomus: Matkalla luovuuden tulevaisuuteen*. Kosmos, 2020.

Kanerva, Jenna; Rönnqvist, Samuel; Kekki, Riina; Salakoski, Tapio & Ginter, Filip. *Template-free Data-to-Text Generation of Finnish Sports News*. University of Turku, 2019. https://aclanthology.org/W19-6125. Retrieved 12.6.2022.

Kauranen, Pekka. *Yle adopts the National Library of Finland's Annif tool for the automated tagging of articles*. Yle, 10.12.2021. https://yle.fi/aihe/a/20-10001817. Retrieved 12.6.2022.

Kjellman, Martin. *"Automation will save journalism" – News automation from the service providers' point of view*. University of Helsinki, 2021. https://helda.helsinki.fi/bitstream/handle/10138/327380/Kjellman_Martin_thesis_2021.pdf. Retrieved 12.6.2022.

Korab, Petr. *Training Neural Networks to Create Text Like a Human*. Medium, 2022. https://towardsdatascience.com/training-neural-networks-to-create-text-like-a-human-23bfdc23c28. Retrieved 12.6.2022.

Kvale, Steinar & Brinkmann, Svend. (2008). *InterViews: Learning the Craft of Qualitative Research Interviewing (2nd ed.)*. Sage Publishing, 2008.

Latar, Noam Lemelshtrich. *Robot Journalism: Can Human Journalism Survive?* World Scientific, 2018.

Lehto, Niko & Sjödin, Mikael. *Automatic text summarization of Swedish news articles*. Linköping University, 2019.

Leppänen, Leo; Munezero, Myriam; Granroth-Wilding, Mark; Toivonen, Hannu. "Data-Driven News Generation for Automated Journalism." In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197. Association for Computational Linguistics, 2017. https://aclanthology.org/W17-3528. Retrieved. 12.6.2022.

Lindén, Carl-Gustav & Tuulonen, Hanna (Eds.), Bäck, Asta; Diakopoulos, Nicholas; Granroth-Wilding, Mark; Haapanen, Lauri; Leppänen, Leo; Melin, Magnus; Moring, Tom; Munezero, Myriam; Sirén-Heikel, Stefanie; Södergård, Caj; Toivonen, Hannu. *News Automation: The rewards, risks and realities of 'machine journalism'*. WAN-IFRA Report, 2019. https://jyx.jyu.fi/handle/123456789/67003. Retrieved 12.6.2022.

Marconi, Francesco. *Newsmakers: Artificial Intelligence and the Future of Journalism*. Columbia University Press, 2019.

McLuhan, Marshall. *Understanding Media: The Extensions of Man*. McGraw Hill, 1964.

Media Audit Finland. *KMT 2021 tulokset julkistettu*. Media Audit Finland, 2021. https://mediaauditfinland.fi/2021/10/13/kmt-2021-tulokset-julkistettu/. Retrieved 12.6.2022.

Merriam-Webster. *The dictionary entry for "medium"*. Merriam-Webster online dictionary. https://www.merriam-webster.com/dictionary/medium. Retrieved 12.6.2022.

Newman, Nic. *Journalism, Media, and Technology Trends and Predictions 2022*. The Reuters Institute for the Study of Journalism, 2022. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-01/Newman%20-%20Trends%20and%20Predictions%202022%20FINAL.pdf. Retrieved 12.6.2022.

Paikkala, Maija. *Tekoälyllä on kyltymätön datan himo ja muut opit STT:n Scoopmatic-projektista*. STT, 2020. https://stt.fi/scoopmaticin-opit/. Retrieved 12.6.2022.

Pollak, Senja; Robnik-Šikonja, Marko; Purver, Matthew; Boggia, Michele; Shekhar, Ravi; Pranjić, Marko; Salmela, Salla; Krustok, Ivar; Paju, Tarmo; Lindén, Carl-Gustav; Leppänen, Leo; Zosa, Elaine; Ulčar, Matej; Freienthal, Linda; Traat, Silver; Cabrera-Diego, Luis Adrián; Martinc, Matej; Lavrač, Nada; Škrlj, Blaž; Žnidaršič, Martin; Pelicon, Andraž; Koloski, Boshko; Podpečan, Vid; Kranjc, Janez; Sheehan, Shane; Boros, Emanuela; Moreno, Jose; Doucet, Antoine & Toivonen, Hannu. "EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions". In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation (EACL2021)*, 2021. https://aclanthology.org/2021.hackashop-1.14.pdf. Retrieved 12.6.2022.

Salmela, Salla. *Automaatiota uutistoimistossa – Tapaustutkimus STT:n toimittajien näkemyksistä tietojenkäsittelyjournalismista*. University of Jyväskylä, 2021. https://jyx.jyu.fi/handle/123456789/74684. Retrieved 12.6.2022.

Samoili, Sofia; Lopéz Cobo, Montserrat; Gómez, Emilia; De Prato, Giuditta; Martínez-Plumed, Fernando & Delipetrev, Blagoj. *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*. Publications Office of the European Union, 2020. https://eprints.ugd.edu.mk/id/eprint/28047. Retrieved 12.6.2022.

STT. *Johto ja omistajat*. Suomen Tietotoimisto. https://stt.fi/omistajat/. Retrieved 12.6.2022.

Vehkoo, Johanna. *Uutisrobotit tulivat vihdoin Suomeen*. Journalisti, 24.3.2017. https://journalisti.fi/artikkelit/2017/03/uutisrobotit-tulivat-vihdoin-suomeen/. Retrieved 12.6.2022.

Yle. *Financial figures 2021*. Yle, 2022. https://yle.fi/aihe/s/10002615. Retrieved 12.6.2022.

Yle. *Lätkä-voitto*. GitHub repository, 2017. https://github.com/Yleisradio/avoin-voitto. Retrieved 12.6.2022.

Zuboff, Shoshana. *Be the Friction – Our Response to the New Lords of the Ring*. Frankfurter Allgemeine, 25.6.2013. https://www.faz.net/aktuell/feuilleton/the-surveillance-paradigm-be-the-friction-our-response-to-the-new-lords-of-the-ring-12241996.html. Retrieved 12.6.2022.

Zuboff, Shoshana. *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, 1988.

APPENDICES

Appendix 1: Pre-Written Questions

Q: What kind of automated tools does [Yle / STT] use for news production?

Q: In what kind of news could automation or AI replace a human journalist? What kind of news is automation or AI better suited to be a tool for journalists to use?

Q: Does [Yle / STT] use templating in producing the basis of news articles? In what kind of articles?

Q: Does [Yle / STT] use machine learning and natural language processing to produce articles?

Q: Does [Yle / STT] use versioning or content summarization, or automatic recommendation of images to accompany an article?

Q: Does [Yle / STT] seek cost savings by automating the news?

Q: When a robot produces the news, who gets the byline? Are such news labeled?

Q: In your opinion, can a piece written by AI go against the guidelines for journalists? For instance surrendering decision-making outside the editorial office?

Q: How do journalists react to the technological development?

Q: How does the audience respond and how do you know it?

Q: What will happen in this field in the next 5 to 10 years?

Appendix 2: Transcript of Interview with Jarkko Ryynänen

Date: 26.11.2021
Place: Microsoft Teams

Sampo Sauri: What kind of automation tools does Yle News use for news production?

Jarkko Ryynänen: That's a kind of a complex question. We actually build all the tools by ourselves, or pretty much. In some parts of the way we might be using some some shopped tools, but anywho, they are pretty much all there homemade.

Like for example for the Voitto robot. We have built everything by ourselves so, homemade stuff, DIY.

Now that you mentioned Voitto, but it's been a while since it's been in the news. It's sort of when it launched there was more about it, but how does Voitto work? And how has it evolved during the past four years?

Actually five years, it was December 2015. Nowadays, Voitto follows politicians. We are gathering information about Members of Parliament. And we make once a month a newsletter that tells what Voitto has done in the past month. But when Voitto was born, he was keen on hockey, NHL to be exact. He wrote ten articles about NHL hockey. And since those days, Voitto has been writing stories about sport. Different kinds of sports: floorball, ice hockey, football and so on.

But pretty soon, 2017 there was a municipal election, so we tried if Voitto would be suitable for political journalism, and he was. So since those days Voitto has been writing a lot about politics. And he will do some stuff in these, they are aluevaalit, it is in English, it's a county election. That's a long way down in the brain, but it it's weird name but it's called the county election.

OK, so Voitto is working for you through this county election.

Yeah.

I read up that there was a bot called Valtteri which was also doing things like municipal elections 2017. Is that a different project altogether or is it sort of combined?

Yeah, it was a different project. We talked about, uh, maybe cooperating, but because of political issues between a government media versus commercial media, we didn't find any any clear land to work together, so to say, so therefore they are separate projects.

OK. I get it. Just to recap, the Voitto robot started out doing sports journalism on ice hockey, then moved on to other sports and politics.

Yeah, that's right.

Alright, good. How would you then say it, is it possible, or in what kind of news could automation or AI replace a human journalist?

At first I have to mention that Voitto is nothing about machine learning or artificial intelligence. Voitto is using premade templates. And you could simplify that Voitto has brief app templates and he fills out data points that those templates. It's, uh, maybe too much simplified, but the basics are like that so you can automate basically anything where you have a machine readable data.

And for those reasons, sports are quite obvious victims. But also, there is pretty much data on politics too, and it's a more interesting area. Nowadays people are talking much about ethics of machines or artificial intelligence and it's not very dangerous to automate sports news because if you make a mistake, it's not so bad, but on a political area, if you make a mistake, it could be a scandal.

So therefore, it's way more interesting to work on a political area, but it's also more dangerous. And then for example about those Parliament Member following things we are gathering information from the Parliament web page. And we are reporting, for example, what they have spoken in the meetings or how they have voted or attended meetings. We are trying to gather all kinds of statistics about Parliament Members' doings and sayings. But for now we haven't gone beyond that.

Human beings, we are pretty good on deducing from data so we can deduce that if you combine these two data points it must mean that there is something happening with this Parliament Member and Voitto is doing nothing like that at the moment. So we are keeping on an almost statistical level so to say.

Yeah, I understand. Are those articles that Voitto does, are they being read by a human journalist before publication? Or is it completely automatic?

At first they were completely automatic and the reason is that if you automate something, it it's a bad idea to make a human being to check if the automation is going smoothly. For example, if you are grinding wheat, it's a dumb idea to put a human being to watch that. The flour will be good because that's the whole point of alternate version, automatization that you don't need those human beings. So we started with that. But nowadays it's more like cooperation between human and machine, so Voitto isn't doing all the articles by or all those newsletters by himself, but they

are like combined work of human beings and machine and therefore you could say that they are read by a human being before they are sent.

Yeah, so they are. The Voitto robot sort of makes an article with all these statistical parts in it, but then a human journalist can sort of add to that.

Yeah, that's right.

It's sort of in a normal journalistic process.

Yeah, that's correct.

My next question then is going on to this AI or machine learning, kind of. Does Yle have machine learning initiatives or are they producing news which leverage those kinds of technologies?

We are using machine learning in recommendation where you have to know what kind of things are interesting to certain kind of users.

And we have been trying like audio to text things and stuff like that, but they are not so widely used. So basically, the recommendation is maybe the most used and we haven't been trying to make synthetic news, so to say that machine wouldn't have any kind of templates, but that algorithm would somehow learn to write human language.

But those are kind of top edge technologies anyway in this planet so therefore we haven't been trying those in Finnish yet, but maybe one day.

Yeah. Thanks for your answer, it's good. It was on my list here of questions, that does Yle use machine learning or like natural language processing to produce text, but maybe in the future.

No, the only owner case that I know in Finnish is Suomen Tietotoimisto case where they tried to produce news with ML algorithm and it was an interesting case, but they ran out of time and ran out of learning materials, so they didn't quite reach the goal in that case, but it's interesting case in any way.

Very interesting, and I shouldn't think that STT are running out of material to train it with.

Well, the basic problem in machine learning is that you need so big pile of learning material that it's unbelievable. It's usually in every case where I have been involved, we have run out of that material. So you need like millions of items.

And for example this, they tried to make a sport news and there is simply not enough sports news in Finnish to learn

that kind of algorithm correctly, at least at those days, there is some gimmicks that you can use in that learning material, but they can, how would I say, affect the end result.

Yeah, I understand. Then does, or how much does Yle use, if at all, automatic versioning of articles or content summarization?

Ah, you mean, like, giving a different kind of article to different kind of users.

Yeah, but yes, that an article would be made into a different kind of, just made shorter or made into a listicle or a maybe sort of different type of article.

I don't know any case that we are using that. I know the BBC is doing that at the moment and as far as I know in Finland we are, you know, like in front page level in that kind of thinking at the moment. Like for example, both afternoon paper media sites, they are different each user so they adapt the content in front page, but I don't know if they adjust that content in article level at the moment. But I know that BBC has done that kind of thing so.

Yeah, I think the AP has been doing sort of content summarization with sort of a machine learning algorithm to make it.

Yeah, that's the different story. We have been trying summarization. But at the moment you can help a human being to make a summarization, but they are not so good quality yet that you could publish them directly to the end user. So they are more like tool kind of things like helpers for journalists.

I'll backtrack a bit. I'm looking more at the tools for journalists and not so much on the recommendation part of how to publish. You said that it can be used as a tool for journalists, but have you sort of played with summarization?

Yeah, we we haven't tried it actually in printing. Many like proof of concept cases. But I don't think that's we are using it day by day basis. So they are more like “hey, let's try out, could we summarize? They just kind of content types and maybe automatically transfer them to audio or something like that.

Yeah, I understand.

Do you have a set of other automatic tools to help journalists, like recommending images to use or videos to use in an article based on the content? So by keywords or?

I'm not sure if I'm the right person to know about that because I haven't been working on those CMS level things in

such a long time, but no that I know of, but I'm not sure. But we are, well, this is kind of basic tools, but when are you having metadata so we have a optimization recommending or saying that these keywords might be good for this article, so that kind of stuff we have been doing years and years.

Yeah.

But I'm not sure if it's any interesting.

Yes, it is a bit interesting that too. What kind of methods do you have to that. And what ontologies or?

Actually, we used Leiki before, but now we are using our in-house product for that.

OK.

And this is more like ethical question. This why we are moving to homemade stuff is because even though firms are selling these things as services, but usually they are kind of black boxes that you don't know how they are working. If there is an error, how you can fix it. But when you do it yourself, you know exactly how it was done and if it's working incorrectly, you know who is going to fix that. Therefore, we are moving more and more to this DIY thing.

That's kind of an interesting point, that you mentioned about these commercial systems often being black boxes, so you can't really know what kind of an algorithm produces the end result. So there's a matter of trust in that thing.

Yeah, that's right.

Maybe you can answer this, maybe not, but when there's recommendation of images that could work with the article, do you know if you make sort of automatic charts like infographics or automatic maps based on data or templates? Do you have like tools where you just input some fields?

Yeah, actually, that's pretty much what what Voitto is doing.

For example, whenever there is a vote on a Parliament, Finnish Parliament Voitto draws a picture of that voting, and if journalists want to use that image, they can use it so. And they are voting every day and so many votes. So there is a thousands and thousands of pictures of votes or voting in Finnish Parliament. And we have been thinking that we should change that algorithm that you could kind of order Voitto to draw picture so he would draw pictures only when it's needed, but it's more simpler to build that kind of machine that just draws picture every time there is a voting. So which test the locking lot of those kind of pictures in our CMS and only few of them will be used in articles, but that's one example that where we are using. And

they are same kind of templates where is the end result and the name of the voting that has been done in Parliament.

Yeah, but now that you went back to mentioning the templates, I have a follow up question there. Can you recap in what kind of articles is templating used? Is it just sports and politics or are you looking at having more different kinds of articles where it's used?

Yeah, we are using them pretty much in every aspects of journalism. Voitto isn't necessarily doing them, but we are using actually both software, I can't come up with the name at the moment, but maybe it comes later, but where you can for example make a bar chart or line chart with some data, that will work in in about every area of journalism so yes, we are using them pretty much everywhere.

Can you explain in more detail, how does the templating work?

In Voitto case, Voitto, traffic, everything. So you put in data. You receive a image. Because we made it that way, but in that software that we are using a well, if they are kind of which sets. Rather than a images. So you put data in inside of widgets and widget is showing it like in bar chart or lunch or or pie chart or whatever shape you would like to use.

And how about the templating for the articles themselves that Voitto is doing? Can you shed some more light on how exactly that's done?

In sports area we are using like they all like sentence templating, that if the game match was a tie at the last minute of the game, so we can say that it was a tight game.

And there was a surprise in the end when other team made a goal in a last minute, so we have this if … else kind of stuff there and we are running through that data and pick up lines for the article, deduced by the data. And the complexity of that is that you have to know before what kind of things would be interesting in certain kind of cases, and there is a danger that you are forgetting some kind of random case that there is no sentence for that kind of situations at all, and we human beings are very good at pointing out those kind of things that Voitto is ignorant, he doesn't know anything interesting in this data set and human being sees right at the first second that that was the interesting part, but because we didn't teach Voitto to recognize those cases, he will not recognize those cases.

So that's the problem with template-based deducing that you have to know beforehand every case that you want that machine to know. And machine learning is way more flexible on those kind of situations. So that's the big benefit of machine learning or artificial intelligence.

So technically, how is it done? Is it a Python script or what kind of a program is it?

Ah, it's so old that I can check that the tech for you, but I can't remember it right away. It might have been Scala, but I'm not sure.

Well, they did so. Series of if and if else statements that pick the right kind of text file.

And actually we have, if you want to see the code, we have Voitto as an open source version. So you can find it.

Oh, you have it on GitHub or somewhere.

Yeah, it's in GitHub. So if you want to see how it how it's works. So you can see it.

OK, good to know. Alright.

And actually Voitto is used, Suomen Tietotoimisto has an own version of Voitto and Ilkka-Pohjalainen also has a version of Voitto.

Good to know.

So we have given it away also.

Yeah, if it's open source then anyone can use it if they want.

Yeah, and we have given them a newer version.

Alright.

Because that open source Voitto is pretty old.

OK. Alright, I'll move on. I have a few questions left about different kinds of things. How have you found that that journalist to react to these kinds of sort of technological features or automatization of news production?

The reactions are pretty interesting. When we started this thing in 2016 and Voitto wrote ten articles of NHL games. And we were so proud when we went to the sports department to show that what we have done. And all the journalists were afraid of what kind of monster this Voitto is and the first or second question was that will this robot take my job? So that was their major issue. They weren't impressed at first that what this kind of automatic thing could do.

And we thought that thing that because every time we went and spoke of this project people were asking that is this some kind of scheme to reduce people in offices, and we try

to say that no, this is so simple machine that how can we ask that? But even though people ask for that. And then we had a name and image in those first articles, we had a name Voitto and image. But for some reason we decided to make a doll of Voitto, something like maybe a little bit over 1 meter height. And since we have this doll, handmade doll and well, we went to speak about Voitto to people started to think that it's actually the doll that makes those stories, and since that doll was kind of cute, so they weren't afraid of that robot anymore, that's very strange thing, and it was a lucky that we decided to make that doll. I can't remember why we, maybe it was because we have to have something to show because it's kind of abstract to show some kind of lines of codes. But it was lucky accident.

So people, instead of being afraid of that robot, they wanted to take selfies with that doll. It was very lucky accident and then we got to speak about what this robot should do next, because people weren't afraid anymore and they started to think that "I have so much dull work that that thing could do and I could do something more interesting."

It was weird but very interesting.

Sounds like a fun experience.

Well, if that's the journalists' reaction, well, how has the audience reacted to these Voitto articles.

Well, it's kind of biased, or maybe a binary reaction. Some people are liking those articles very much in those articles that were made solely by Voitto, some people hate Voitto very much and still do. They actually said that they won't read anything from Yle anymore because there is a robot doing human's job. But since in sports area, some people are loving, for example teletext pages. And these are kind of, well, teletext pages but maybe version 2.0. There is more stuff in that, but basically the same thing that it's simplified story of a match, how it went, what was the end result. But anyway, they were read, the articles. But then we had a problem that a human journalist wanted to work on those sports area that are popular, so therefore Voitto had to write stories of those sports that are not so read. So he was doing more like niche articles.

Yeah, human journalists wanted to do NHL.

Yeah, that's right, that's actually the case that it was the last time that Voitto did NHL game stories that December night on 2016. Because human journalists wanted to write those stories.

OK, that's interesting.

But nowadays when they are cooperation between human beings and robot, they are, we are making them more rarely. So we

are not publishing like 1000 articles per day, but maybe a couple of articles per month and they are kind of popular, so plenty of people are reading those articles. But it's hard to say what part of that is due to the robot, and what part is due to the people. So it's cooperation.

Well. When you mentioned the text, the Voitto logo and name on the article. So how is it now? When Voitto or another robot produce news, who gets the byline?

Voitto gets own byline and those human beings involved get their own so everyone is mentioned. So Voitto is like one of the colleagues of journalists, so they are in the same level.

Everyone is mentioned. From a part of that kind of news where Voitto is involved. Are they labeled in any other way like is there a text at the end explaining Voitto anymore?

We are not using in the end, but we are saying that for example, if there is a widget inside of the article, so we are telling that this image or widget was made by a robo journalist called Voitto.

OK.

For example, those Parliament voting images. Let's say them for example. So whenever we are using them we are telling that this was made by Voitto and we were using Parliament data.

Well then, this is an opinion question but and could have a kind of hypothetical when this kind of news are not written, but in your opinion, can a piece written by AI go against the guidelines for journalists?

Sure they can. Yeah, yes they can.

Yeah.

But yeah. I don't recommend it, but it's possible, of course. And therefore that's one of the reasons that we are using templates at the moment. Because, in the end, there's always a human being who is responsible of those articles. And usually those human beings want to know how this machine works. So when you're using templates, you can say that if we put in this kind of data, we will get this kind of output. But then machine learning, it's more complicated. You can say that if we put this kind of data in, we get pretty much something like this, but we can't be sure. And usually that's not the answer that human being responsible wants to hear.

So, the more you automate, the less you have a control in detailed level. But it's the same problem with human beings. Human journalists can go amok, so what if they write something that's illegal? You can tell them to stop. We can pull the plug on the robot so. In the end, you have to trust something. But for some reason we do not trust machine as much as we trust human beings. And maybe there is a good reason for that. Usually people are more afraid of machine, doing bad things than human beings doing bad things. And there is one point that machines, they are usually pretty fast, so if they're doing bad things, they are so fast that they can do so many bad things in such a short time period so there is a point in that way.

How about some of the guidelines? There's this surrendering of decision-making outside the editorial office. Would you still consider AI a part of the editorial office or not?

Well, there there you have the the case of black boxes and the in-house boxes, so therefore it's in my opinion it's very important to do those boxes by yourself because then you know how they work and then you don't give that that power or authority to the outer entities, but since we already are using black boxes in this planet and in Finland, so, if you are very strict for example afternoon papers are breaking that rule already. I'm quite sure that they haven't built those recommendation softwares by themselves. But it's weird rule anyway, because how do you draw the line between journalistic office and those. Is a freelancer part of that team or not, so it it's weird role anyway, in my opinion.

Yes. It has been breaked in so many ways in so long time, so I know what they were meaning, but it it's weird how that how it's written in those those rules.

Yeah. Gotcha. OK, we're nearing the end of my premade questions.

How do you think this field will evolve in the next five years?

Well, we will have that synthetic text producing in Finnish in five years, I'm quite sure of that. I'm not sure if, for example, in that case, is it GTP or GPT 3? Uh, it can write English. But that's kind of semi-intelligence, but usually the end result is so easily nonsense. So the next problem is how to make it non-nonsense? And that's actually the problem with that Suomen Tietotoimisto case. That they got text machine could write a lot of text. But to tell a coherent story with that text is another issue. So I think that in five years, it's a long time, so surely we have the Finnish version of that. But maybe we have a English version that provides sensible text. Finally, maybe first Pulitzer will be given to the machine. And nowadays, actually I saw a couple of cases where artificial intelligence or machine learning made videos by text. So you write some

kind of story and it's transferred to video or animation. So in in a rough you could make animations by doing a script and running it through this kind of machine learning thing and wow, you have a finalized animation.

Pretty cool.

Yeah. And in the same time, pretty horrible because some days those machines will be so good that no human animators will be needed anymore. And there we get that question that we got where we started with Voitto that will this machine take my job.

Well. Yeah, I mean, you're completely right. And when you been looking at these AI made faces and now sort of Unreal Engine animating with those faces that look completely like humans, it won't take long.

And it's interesting that they are. Uh, going for the television and film industry so they are? They are no not anymore solely on their game genre, but they are moving strong with the film industry and television so.

Yep.

It's interesting.

Right, that are already used by Lucasfilm.

Yeah, that's right.

Well, I hope you've enjoyed being interviewed, but my last question is a counter question. Do you have any thing that I haven't sort of mentioned that you would like to tell me about, what you do or what Yle does in this field that should be included in a thesis like this?

Not that comes to the mind, but I recommend if you have time and energy. For example, we have a Nordic country meeting thinking about the ethics of artificial intelligence.

So, it's a hot topic at the moment as you know.
And there is very interesting things happening in that area. Because the ethics of artificial intelligence are basically the ethics of human beings, and since we are so different, we do not have a certain rule book that everyone could follow. For example, in Sweden there is robots writing stories as you mentioned about sales of houses and tenants. And in those articles the Swedish publish the price of the house, the name of the owner or who bought it and the age of who bought it. And we had one of these meetings where there's the makers of that those were asking us from different Nordic countries: “But how do you see this? Is it OK to make these kind of stories? Is it OK to say that this house was sold for this amount.” And pretty much everyone

raised their hand that this is OK, and then the next question was that “Is it OK to publish the name of the new owner?” And for example, Swedes and Danish people were like, “Yeah it's OK we are doing this all the time” and we Finns were like “What the hell? Who would publish this kind of thing?”

And the first question that is it OK to publish the age of the new owner? And we Finns are like that “Yeah, sure, it's statistical data, so why not?” And the Swedes and Danes were like “No way, that's too much. We draw the line there”, so there you see that even though we are kind of close with Swedes and Danes, we have so big differences in those ethical areas, so how can we teach machine to follow certain kind of ethics when we can't even decide it by small amount of people that what is right? What is wrong? So that's very interesting area that's so interesting, super interesting.

Appendix 3: Transcript of Interview with Salla Salmela

Date: 20.12.2021
Place: Microsoft Teams

Sampo Sauri: Sä oot STT:llä töissä. Mitä sä teet siellä?

Salla Salmela: Mä oon meidän 24h-toimituksen tuottaja. Eli siis käytännössä vastaan meidän uutistuotteista ja se tarkoittaa taas sitten tämmöisiä työn kulullisia ja vuorokohtaisia ohjeistuksia. Sitten mä päivystän meillä siis ihan tässä meidän esihenkilöringissä kaikkia sairastumisia sun muita. Ja sit mulla toinen puoli mun työajasta on tuollainen mediapalveluissa eli mulla on sitten toinen jalka myös tuolla asiakastöissä, eli olen meidän media-asiakkaisiin yhteydessä sitten siitä että mitä he toivoo STT:n uutisoinnista tai jostain jotain muita palveluitamme mitä meillä on niin mitä he niiltä toivoo eli vähän tämmöinen kaksi hattua päässä: Toinen on siellä journalismin puolella ja sitten toinen on tuolla asiakastöissä ja mun tausta on ihan siis olen toimittaja taustaltani että ennen kun tein tätä niin tein itseasiassa meidän yövuoroja Australiassa. Eli tuota minä olen nyt nelisen vuotta varmaan ollut tässä tuottajana ja sitä ennen olin siellä Sydneyssä uutistoimittajana ja sitten sitä ennen on ollut erinäisissä uutishommissa Suomessa.

Mä voisin aloittaa ensimmäisellä kysymyksellä, minkälaisia tämmöisiä uutisten automointityökaluja STT käyttää uutistuotannossa?

No se vastaus riippuu vähän ehkä siitä, että mikä katsotaan sellaiseksi uutisten automatisoinniksi. Onhan ihan joku oikoluku on tavallaan automatisointia, mutta jos mietitään näitä tämmöisiä vähän sitä kekseliäämpiä, niin meillä on muun muassa meidän hälytyspalvelut on osin automatisoitu, eli kun meillä on siis tämmöinen yöpalvelu jota myydään media-asiakkaille, että jos asiakkaalla itsellään ei ole yöpäivystystä niin sitten STT hälyttää sen media-asiakkaan ikään kuin niin kuin töihin siinä vaiheessa, jos tapahtuu jotain isoa vaikka joku tämmöinen iso luonnonmullistus ulkomailla yöaikaan niin sitten tehdään tämmöinen niin sanottu valtakunnan hälytys, niin siinä on se soittorumba automatisoitu niin, että toimittajan tarvitsee vain yhden kerran kirjoittaa tekstiä, että mitä me halutaan sinne hälytyksestä sanoa ja sitten sellainen automaatiojärjestelmä alkaa soittaa sitten meidän asiakkaita / päättäjiä läpi ja herättää heidät. Niin se on yks esimerkki. Sitten meillä on tällainen niin sanottu Pikkulintu, joka on siis toimittajille tällainen selaimen lisäosa joka on meillä tuota semmoisessa projektissa kehitetty, minkä avulla pystyy seuraamaan verkkosivuja, että joko siis ihan silleen, että se voi säätää jonnekin verkkosivulle, että ”ilmoita minulle, kun tämä sivu päivittyy” tai sitten sen voi säätää sinne verkkosivuille että ”ilmoita minulle jos tällä sivulla tai

tämän sivun osiossa A, B ja C koskaan mainitaan vaikka Suomi tai joku luku päivittyy tai joku asiasana ilmestyy sinne sivulle”, niin se on tarkoitettu siihen, että kun meillä tehdään siis tosi paljon tällaista nopeata päivän päällistä reagoivaa uutistyötä, niin siinä se tiedon etsintä on sellaista, että koko ajan pitää olla sekä koneessa että aivoissa monta tabia auki niin vähän vähennetään sitä semmoista sinkoilun määrää että Pikkulintu kertoo sitten se visertää sieltä jos jossain tietyssä jossain tietyllä sivulla vaikka päivittyy joku odotettu asia tai odottamaton asia. Me ollaan käytetty sitä muun muassa noihin kansalaisaloitteiden seurantaan, että sitten kun me tiedetään, että joku aloite luultavasti kohta menee sitten pöydälle niin voidaan säätää sinne Pikkulintu sitten.

No meillä on Tarkkailija joka käyttää hyväkseen tuota meidän tytäryhtiön Retrieverin tämmöistä mediaseuranta-osumadataa niin sitä me käytetään sitten siihen, että me katsotaan meidän tuotannon läpimenoa meidän asiakkaille. Eli kun me ollaan uutistoimisto, niin meillä ei ole suoraa pääsyä tällaiseen julkaisun dataan mitä julkaisijoilla on eli jos nyt vaikka Savon Sanomat julkaisee printissä ja verkossa STT:n jutun niin Savon Sanomat saa sen suoran pääsyn tavallaan heidän lukijatietoihin esimerkiksi siihen, että kuinka monta kertaa sitä juttua on klikattu netissä ja kuinka kauan sitä luettu. Me ei sitä tiedetä kun me ei olla julkaisija, mutta sitten meillä on tämä Tarkkailija-työkalu, jonka avulla me pystytään näkemään sitten ainakin se, että mihin printteihin joku juttu on mennyt läpi ja mihin verkkosivuille se juttu on mennyt läpi ja esiintyykö se mahdollisesti myös sitten jossain muualla verkossa kokonaisena, että periaatteessa se löytää myös tällaiset väärät käytöt, mutta siihen me ei oikeastaan niinkään sitä käytetä, vaan siihen, että katsotaan, että mitä meidän juttuja asiakkaat käyttää paljon, mikä tarkoittaa siis sitä, että on tehty jotain oikein ja sitten että löytyykö sieltä jotain vaikka toistuvia juttutyyppejä mitä meidän asiakkaat eivät ole ollenkaan, mikä tarkoittaa sitä, että me käytetään meidän resurssit johonkin sellaiseen mikä ei ole hyödyksi meidän asiakkaille että parempi lopettaa sellainen. Niin tähän me käytetään Tarkkailijaa ja sitten tuota sieltä toki pystyy yksittäinen toimittaja katsoo myös omien juttujensa tuota menestymistä eli jos haluaa sitten ihan silleen juttu jutulta.

Sitten meillä on ollut ollut tuota kehityksessä ja kokeilussa tekstigeneraatiota myös siis hyvin tällaisilla, oikeastikin keinoälyllisillä komponenteilla, että yksi tuon Turun yliopiston kanssa Google-rahalla kehitettiin tämmöinen Scoopmatic, joka on tämmöinen itseoppiva kielimalli. Se opetteli ensin siis meidän arkiston digiarkiston perusteella kirjoittamaan suomea, kirjoittamaan STT-tyyliä, kirjoittamaan uutisia ja sitten siitä silloin ruvettiin antamaan vähän parametreja että mistä pitäisi kertoa eli pyrittiin tekemään siitä sellainen malli, joka olisi pystynyt uutisoimaan jääkiekosta. Eli sille olisi annettu aina niin kun otteludata, tuore otteludata ja sitten kaiken sen

perusteella mitä se olisi tiennyt meidän aikaisemmista jääkiekkouutisista, niin se olisi kirjoittanut sitten tekstin.

Ja se pystyi siihen. Kyllä sen suomi oli varsin hyvääkin, mutta sitten se ongelma mikä siinä oli niin oli just tää mistä Leppäsen Leo varmaan puhui siellä missä olit eli se halusi noi että sitten se saattaa yhtäkkiä keksiä sinne vaikka jonkun semmoisen joukkueen, joka ei ole siellä jäällä ollutkaan. Tai sitten se saattaa napata sellaista tavallaan kieltä mitä ei käytettäisi, että se ei ymmärrä tämmöisiä ikään kuin tilanteita, eli se saattaa vaikka sanoa että joku joutui suihkuun kun se vaan tietää että usein tällaisissa tilanteissa puhutaan jääkiekon yhteydessä tästä. Mut et sit oikeasti kukaan ei ole komennettu pois sieltä kentältä.

Niin se oli tosi mielenkiintoista nähdä, että suht pienelläkin semmoisella opetusaineistolla niin nää oppii sitten näinkin pientä kieltä kuin suomi ja siis että se tuotanto mitä ne tekee on ikään kuin mielenkiintoisempaa lukijalle kyllä kuin tämmöinen toistuva template, mutta sitten kun puhutaan journalismista, missä faktat pitää olla just eikä melkein, niin ei vielä päästy sellaiseen pisteeseen, että se olisi ollut niin niin kuin faktapohjaista se tuota koneen suoritus, että me oltaisiin voitu vaikka suoraan julkaista jotain.

Tai sitten että se olisi ollut niin faktapohjaista, että toimittajan ei olisi tarvinnut käydä jokaista lyhyttä tekstiä erikseen läpi ja sitten sieltä ikään kuin toimitussihteerinä sanoo et ok, niin silloin tullaan sitten siihen, että jos meillä on tällainen solution niin se ei sitten kuitenkaan säästä meiltä sitten resursseja ja aikaa jota me halutaan että se säästäisi. Mutta tämä oli mielenkiintoinen kokeilu ja ehkä siitä sitten joskus vielä voi jotain poikia lisää.

Onko se Scoopmatic nyt edelleen käynnissä tai onko tämä niin kuin hyllyllä vai?

Se projekti on päättynyt, mutta toki malli on olemassa.

Hyvä. Nää on kaikki kiinnostavia. Tavallaan automaatio on ehkä vähän liian semmoinen kattotermi sitten niin kuin tähän, mutta nimenomaan se tekstin generaatio on tämän työn fokuksessa, eikä sen ympärillä olevat asiat niin kuin esimerkiksi julkaisu tai suosittelualgoritmit tai tämmöiset, niin ne jää niin kuin mun tämän työn skoupin ulkopuolelle eli tämmöisistä mielellään kuulen vielä lisää. Toi voi siis kuulostaa tosi mielenkiintoiselta jutulta toi Scoopmatic. Mistä sä luulet, että se jäi kiinni? Oliko siis tätä koulutusaineistoa liian vähän?

Siis ei varmaan pelkästään siitä. Tähän osaisi paremmin vastata se Turun yliopiston tutkijaryhmä, joka loi sen mallin, koska sitten mennään sellaisiin matemaattisiin

algoritmeihin, että meikäläinen ei todella ymmärrä, mutta siis siinä hekin sanoi sitä, että sitten kun mennään tommoiseen tuon tyyppiseen kielimalliin joka oppii itse niin ikään kuin sen päätöksenteosta tulee sellainen musta laatikko ja siis käytti tätä vertausta että sinne ei näe että se tekee jonkun päätöksen, mutta se ei koskaan selitä ihmiselle että miksi.

Sitten jos esimerkiksi tämmöinen tyhmä template-kone tekee virheen niin sitten vaan mennään ja katsotaan minkä templaten siis ikään kuin siinä koodissa on se virhe koska se toistuu siellä jossain määrin loogisesti ja se voidaan jäljittää. Mutta se että jos Scoopmatic tekee tekstin, siis todellinen esimerkki, jossa tuota ensin alkaa jääkiekkopeli mutta sitten lopussa kaikki vaan ovat vakavasti kuolleina siellä jäällä. [Laughs] Mitä me ollaan? Sitä on vaikea tavallaan kysymys siltä koneelta, että miksi valitsit näin? Minun oletukseni oli, että se saattaa siis liittyä siihen, että sitten jos meidän koko arkistoa katsoo, niin uutisethan usein kertoo konflikteista ja sitten taas konfliktien ja urheilun kieli on vähän samanlaista. Että olisiko se sieltä sitten jotenkin niitä poiminut, mutta että se on ehkä sen mallin luonteessa siinä, että se halusi nuo kuin siinä, että jos sinne laitettaisiin hirveästi ja hirveästi lisää tekstiä että sitten sinne melkein pitäisi laittaa jotain templaten tapaisia asioita raksuttamaan sinne taustalle, jotka sitten vähän hillitsi sitä sen hurjaa mielikuvitusta.

Joo tosta se mun mielestä Leo puhui siellä Embeddiassa kai kanssa, että kun voi rakentaa näitä järjestelmiä vaan tiettyihin osiin käytetään neuroverkoilla tehtyjä.

On tosi mielenkiintoista kuulla myös noita tommoisia ikään kuin oikeita sattumuksia, että minkälaista se on ollut. Onko niitä julkaistu missään niitä ikään kuin esimerkinomaisest niitä lopputuloksia.

Eipä taida. Meillä oli semmoinen yks asiakaspäivä missä näitä esiteltiin kyllä mutta tota.

Ja siellä jossain mun lähdeluettelossa niin on varmaan suora linkki sinne Turun yliopiston ainakin yksi tällaiseen tuota tutkimuspaperiin jonka ne kirjoitti juuri Scoopmaticista jos haluaa tietää siis siitä että miten se toimii niin tuota sieltä löytyy enemmän että mä en tosiaan. STT tavallaan antoi siihen sen sellaisen journalistisen kehikon, että just evaluoitiin sitä että mitä miten se kielimalli toimisi journalismissa ja tavallaan kerrottiin mitä me halutaan, mutta sitten kaikki se semmoinen korkeampi matikka tehtiin siellä yliopistolla.

Joo, ymmärrän. Tosi hyviä vastauksia. Entä minkäslaista, tämmöistä automaattisen uutistoiminnan tiimoilta, minkälaista yhteistyötä STT tekee muiden uutisorganisaatioiden kanssa? Jos teette.

No siis esimerkiksi Ylen kanssa on ollut yhteistyötä aikaisemmin, kun heillä on tämä Voitto-malli niin tuota hmm. Heiltä on sen Voiton avulla sitten saatu tämmöisiä just jääkiekkotekstejä, niillä sitten testattiin sitä meidän asiakkaiden tavallaan kiinnostusta ja innostusta sen tyyppiseen tekstiin ja sen tyyppiseen julkaisemiseen. Tällä hetkellä se, että meillä olisi tekstin tuotannon kanssa yhteistyötä muiden mediatalojen kanssa niin eipä juuri oikeastaan ole. Ne on sitten enemmän niitä sellaisia keskusteluja, että paitsi että me keskustellaan meidän asiakkaiden kanssa, niin me keskustellaan siis muiden uutistoimistojen kanssa, eli on tämmöinen Minds-verkosto, semmoinen kansainvälinen verkosto, johon me kuulutaan ja siellä vaihdetaan paljon ajatuksia.

Siinä tuota tekstin tuotannossa ehkä se ongelma on tavallaan se, että siinä olisi paljon potentiaalia ja paljon sellaista, että jos sen pystyisi toteuttamaan niin se olisi hienoa ja me tiedetään, että esimerkiksi meidän asiakkaalla olisi kiinnostusta tietynlaiseen valmiiseen automatisoituun tekstiin.

Mutta sitten se kuitenkin se, että sellaista saadaan tehtyä vaatii tietynlaisia resursseja ja sitä taas sitten harvalla uutistoimituksella on. Eli tavallaan olisi hirveän hienoa jos me pystyttäisiin esimerkiksi tuottamaan urheilutuloksia entistä enemmän tämmöisenä automaationa. Ja siis semmoisia projekteja meillä on ollutkin ja sitä mietitään, että miten se onnistuisi. Tai sitten vaikka siis jotain just tällaisia template-tyylisiä tekstejä vaikka jostain suosituimpien lajien alasarjoista tai jostain maakunnallisista sarjoista tai jostain siis sellaisista peleistä, mistä me ei voida tehdä juttuja tällä hetkellä sen takia, että volyymi on iso ja meitä on vähän. Eli että voitaisiin tarjota jotain sellaista mitä me ei ihmisvoimin voida tehdä, niin se olisi yks semmoinen potentiaalinen juttu. Tai sitten että voitaisiin tarjota nopeammin jotain sellaista mitä tehdään tällä hetkellä ihmisvoimin. Siis esimerkiksi niin että jos oltaisiin semmoisessa blue skies -mallissa niin että Scoopmatic olisi pystynyt kirjoittamaan sitten jokaisesta jääkiekko-ottelusta semmoisen hyvin lyhyen sähkemuotoisen juttupohjan siinä vaiheessa, kun se data tulee sinne siitä pelistä ja sitten toimittaja olisi vaan katsonut sen läpi ja täydentänyt, jolloin sitten taas voitetaan minuutteja mikä sekin on meille tärkeätä. Eli joko niin, että voidaan tehdä jotain sellaista kokonaan uutta mitä ei pystytä tekemään ihmisvoimin, mikä olisi kuitenkin ihan mielenkiintoista tai sitten että se voidaan tehdä nopeammin.

Mutta templateissa on se huono puoli, että siihen tarvitaan aina se ihminen, joka ne ohjelmoi ja tuoreuttaa ja joka niitä ymmärtää ja jos ei sitä ole siis toimituksessa valmiina niin että mistä se resurssi kouluttaa? Mistä se tavallaan resurssi sille ihmiselle tehdä otona sitten nämä templatet ja sitten niissä muissa tuota taas näissä malleissa on se huono puoli sitten kun miettii siis ihan tekstin tuotantoa, että jos ne hallusinoi, niin silloinhan

esimerkiksi STT ei voi niitä käyttää koska meillä se luotettavuus on ihan ensimmäinen arvo. Jos meidän uutisiin ei voi luottaa, niin eihän meillä sitten ole mitään, että ei me semmoista parasta arvausta laiteta ulos.

Just näin. Tässä on puhuttu paljon urheilu-uutisista, niin onko mitään muunlaisia uutisia mihin te olette kokeilleet tämmöisiä tekniikoita, malleja?

No siis tällä hetkellä se mitä esimerkiksi Helsingin yliopisto on tuossa Embeddiassa miettinyt, niin kun siis sitten niitä aihealueitahan rajaa myös se, että mistä on olemassa semmoista strukturoitua hyvää dataa. No urheilusta on olemassa ja se on toisteista, ne tapahtumat toistuu niin kun niitä tarvitaan isolla volyymilla sellaisia tekstejä sen takia urheilu on siis yks mitä me ollaan, mihin me ollaan kiinnitetty huomiota. Mutta mitä esimerkiksi toi Helsingin yliopisto nyt on työstänyt niin on koronadata, koska sitten tämä pandemia tuli vähän silleen pyytämättä ja yllättäen ja kaikessa muussa on varmasti huono, mutta paljon siitä olisi strukturoitua dataa eli että pystyisi vaikka sitten päivittäin tekemään jonkun tämmöisen raportin vaikka Suomen tilanteesta verrattuna johonkin Euroopan tilanteeseen tai kaikista maista erikseen tai niistä maista minkä toimittaja tilaa, tai jostain niin kuin kertoo koneelle, että siellä missä on eniten muutoksia niin se luultavasti on se kiinnostavin, kirjoita siitä. Että tällainen pandemia on yksi esimerkki. Sitten siis ihan tällaiset niin kuin sää, liikenne, talous. Taloudessa niitä käytetäänkin jonkun verran maailmalla, että uutistoimistoille muistaakseni olisiko AFP:llä on semmoinen joku malli joka minun käsittääkseni on kyllä template eikä siis mikään AI mutta siis se kirjoittaa näistä tämmöisistä pörssikatsauksista. Ja siis tavallaan tällaisista toisteisista niin sanotusti toimittajalle vähän tylsistä tehtävistä, niin se tekee sitten sen ja sitten toimittaja voi suoraan tarttua puhelimeen ja soittaa toimarille, että ”miten teillä nyt näin hyvin yhtäkkiä menikin?” eli tavallaan pääsee sitten hyppäämään yli sen semmoisen ihan kaikkein turruttavimman homman. Ja kun me ollaan uutistoimisto meille on ulkoistettu paljon sellaista uutisointia mitä mediat haluaa sen seurata, ei välttämättä halua sitä tehdä omilla resursseilla. Esimerkiksi joku tällainen onnettomuuksien seuranta, niin me tehdään sitä paljon meidän asiakkaille. Ja tavallaan siis myös tämmöiset onnettomuus uutiset olisi jotain mihin pystyisi miettimään tekstigeneraatiota, mutta siinä sitten taas se data mitä on olemassa. Se ei ole strukturoitua, vaan se saattaa tulla pelastuslaitokselta aina vähän missä muodossa nyt tulee, että kuka sattuu olemaan tiedotusvastuussa.

Me puhuttiin noista termeistä pikkaisen kanssa ennen kuin mä painoin nauhoituksen käyntiin, mutta minkälaisia termejä STT käyttää tämmöiseen?

Me ollaan sisäisesti puhuttu siis robotiikasta paljon, sitten taas toisaalta me kyllä ymmärretään, että se ei ole välttämättä aina joka asiassa paras termi. Siitä tulee

joskus vähän vääränlaisia mielikuvia, usein vielä sitten semmoisia vähän ehkä utopistisiakin mielikuvia, niin sitten me ollaan myös käytetty ja puhuttu aika paljon automatiikasta ja automatisoinnista, ja sitten sisäisesti keinoälystä silloin kun kyse on keinoälystä. Mutta että kyllä nää on tietyllä lailla vähän sellaisia, on vain vaikeata valita sitä termiä, että mitä käytetään, koska tuntuu siis silleen, että niitä muutenkin alalla käytetään aika villisti ja sikin sokin niin ei ole oikein semmoista yhtä ainoata oikeata, mutta robotiikasta meillä on jotenkin ehkä vakiintunut se tällaiseksi, mutta myös automatisoinnista puhutaan.

Joo no termit vaikuttaa siltä että ne vielä ikään kuin hakee paikkaansa, että mistä sitten ala oikeasti tulee puhumaan. Tosi monia eri termejä nuo, ja aika limittäisiä.

Miten sanoisit, miten uutisten automatiikka on kehittynyt viimeisen 5 vuoden aikana? Ja tässä mä puhun edelleen just siitä uutisten tai artikkelin automaattisesta tekstin tuotannosta.

No meillä tietysti kun meillä ei julkaisussa tällä hetkellä ole tuota niin se on vaikea sanoa että miten se on sinänsä kehittynyt. Se sanottaisiinko, että semmoista kiinnostusta siihen on ja tavallaan sellaista näkemystä, että missä se voisi auttaa parhaimmillaan, mutta sitten siihen aina tuntuu jotenkin törmäillä näihin resurssiongelmiin ja sitten toisaalta siihen luotettavuusongelma on ehkä siis noin viidessä vuodessa ainakin se ymmärrys on mun mielestä kehittynyt, siis myös paitsi meillä niin siis alalla, että ymmärretään sekä ne mahdollisuudet, mutta sitten myös ne koneiden rajat, että ehkä enää nyt ei ole ihan sellaista niin samanlaista puhetta, siis nyt en tarkoita STT:llä vaan alalla yleensä, siis siitä että robotit tulee ja vie kaikki meidän työt. Siitä nyt ehkä aletaan ymmärtää jo, että no sitten ei ainakaan jää jäljelle kauhean hyvää journalismia josta kukaan haluaa maksaa [laughs] jos se kokonaan koneelle annetaan että sitten se on ehkä jotain muuta viihdearvollista kuin journalismia.

Joo no mutta tähän tähän jatkaen sitten. Minkälaisissa uutisissa automaatio tai tekoäly voisi korvata ihmistyötä, tai onko tämmöistä?

No ehkä just niissä urheilutuloksissa se pystyisi tavallaan korvaamaan ihmisen silloin kun se toimisi luotettavasti, niin se voisi korvata ihmisen ainakin tämmöisissä ensiversioissa, jotka perustuu täysin johonkin strukturoituun dataan. Mutta sitten siinä vaiheessa kun pitää taustoittaa ja pitää suhteuttaa asioita ja pitää niin kuin tavallaan selittää sitä ympäröivää maailmaa, että tämä muutos kertoo nyt ehkä tästä tai tämä muutos näyttää nyt pahalta, mutta se johtuu siitä, että samaan aikaan meillä on vaikka globaali pandemia, niin sitä on huono antaa koneelle. Sellainen ehkä missä kone saattaa olla parempi kuin ihminen niin se tekee vähemmän tällaisia ns. tyhmiä virheitä eli että se

ei katso jostain isosta lukumassasta, että numerosta kahdeksankymmentäkaksi tulee kaksikymmentäkahdeksan koska aivot on niin väsyneet, että tällaisia se ei tee. Ja siis tällaisessahan se pystyisi auttamaan ihmistä eli tavallaan just käsittelemään sitä dataa nopeammin pistämään sitä nopeammin kuin erilaisiin laatikoihin ja muotoihin, niin kyllä.

Se vähentää typoja.

Mutta sitten heti jos siinä on jotain semmoista journalistista harkintaa enemmän, niin kyllä mun mielestä siinä vaiheessa sen ihmisen pitää edelleen olla läsnä ja kuitenkin toimitushan sen vastuun sitten siitä kantaa, että et sä voi laittaa algoritmia minnekään Julkisen sanan neuvostoon, että tavallaan ihan täysin sitä vastuuta ei voi kyllä ulkoistaa sitten sille koneelle. Ja mehän ollaan sitten mietitty niin, että jos tällaista siis jossain määrin automatisoitua sisältöä sitten julkaistaisiin, niin totta kai se pitäisi sitten kertoa sekä asiakkaalle että yleisölle siinä yhteydessä, että ”tämä jääkiekkouutinen on tuotettu mallilla x ja se perustuu tekniikka x” että tavallaan ihminen sitten tietää mitä hän lukee.

Joo, sä niin ennakoit mun kysymyspatteristoa täällä, että mä vaihdan järjestystä. On mukana nämä Julkisen sanan neuvostot ja kenelle uutinen kirjoitetaan, sanoitkin siitä pikkaisen mutta kysynpä silti tämän kysymyksen uudestaan niin saan sen.

Kun tämmöinen robotiikan avulla tehdään uutinen, niin kenen nimi siinä uutisessa lukee ja miten ne merkitään lukijaa varten? Jos ollenkaan.

No toi on varmasti siis jotain semmoista mitä pitäisi aina miettiä tapauskohtaisesti riippuen, että mikä se tekniikka on. Mutta jos se nyt olisi vaikka tämmöinen NLP-malli joka kirjoittaa vaikka sen ensimmäisen version siitä urheilusta tai taloudesta niin kyllähän siinä sitten pitää olla siis merkintä siitä, että se on automatisoitu tekstiä ja semmoinen lyhyt selitys siitä, että perustuen mihin, mutta että kyllähän siinä silti esimerkiksi meillä niin krediitit olisi kuitenkin STT eli että siis viimeisen vastuun kantaa se talo, joka käyttää sitä mallia tietysti sitten.

Pitää lähteä myös että mistä se koska oletus on nyt. Totta kai sieltä tulee sitten ulkopuolelta esimerkiksi sitä dataa, että mistä se data on otettu, että lähdetään sitten vaikka että Jääkiekkoliitto tai mistä se vaan tulisikin, Ilmatieteen laitos, että tavallaan siinä näkyisi myös sitten se, että mistä se käsitelty data tulee.

Joo STT:n tapauksessa niin merkitäänkö teillä niihin uutisten yhteyteen sen itse toimittajan nimeä? Vai onko se aina STT?

Joo, jossain tapauksissa meillä on siis vaan vakiintunut semmoinen käytäntö. Siis aina tiedetään kyllä, kuka on minkäkin version tehnyt, mutta tuota ihan lukijalle asti, niin meillä yleensä nimi näkyy siinä vaiheessa, kun se on hieman pidempi se juttu. Eli kun me tehdään niin paljon siis semmoista nopeata ensitietouutista, mikä voi siis olla ihan yksi virke pienimmillään, niin niissä ei ole siis toimittajan nimeä. Mutta sitten heti kun se menee semmoiseen pidempään juttuun, varsinkin sitten kun siinä alkaa olla sitä omaa tiedonhankintaa, siinä on haastateltu ihmisiä, siinä on sitaatteja, niin sitten tulee kyllä se toimittajan nimi siihen mukaan myös ihan vaan sen takia, että se liittyy läpinäkyvyyteen. Tietää kuka on kysymyksen kysynyt. Sitten se on tietysti vähän asiakkaalla, ehkä vähän varioi sitten vielä, että miten julkaisijat käyttää niitä meidän toimittajan nimeä et siis kyllähän meidän se STT-credu kyllä näkyy siinä aina, mutta sitten se, että näkyykö aina tuota meidän toimittajan nimi sitten siellä julkaisupäässä niin se vähän varioi, mutta yleensä näkyy kyllä joo.

OK. Joo hyvä. Puhuttiin tuosta Julkisen sanan neuvoston tai journalistin ohjeiden roolista. Mitä mieltä sinä itse olet, voiko tekoälyn tuottama uutinen mennä journalistin ohjeita vastaan? Esimerkiksi journalistisen päätösvallan luovuttamisesta toimituksen ulkopuolelle.

Jos se on huonosti toteutettu, niin kyllähän se vaan voi. Siis tavallaan, että jos niin, että jos sinne nyt vaikka sitten yhdistetään koneeseen, joku ihan huuhaa datalähde ilman että sitä tarkistetaan että mistä tulee ja mihin se perustuu ja onko se oikeasti validia ja sitten vaan pistetään tavallaan julkaisu päälle ja jätetään se siihen oman onnensa nojaan niin kyllähän sitä silti sitten tulee harjoittaneeksi joukkoviestintää paitsi että se on siltä harjoittajalta valvomatonta ja mahdollisesti tosi huonolaatuista et voi että voihan sieltä sitten siis eihän se kone sitä tiedä jos se vaikka siellä tehtailee jonkun, mä en osaa keksiä tavallaan sitä esimerkkiä, että mistä voisi tulla vaikka kunnianloukkaus.

Mutta siis silleen että konehan sitä ei huomaa, että kyllä ihmisen pitää sitten jossain määrin pitää silmällä.

Sitten on tietty eri asia, että voihan siis tulla jotain niin että on ihan täysin validi datalähde, vaikka sitten sinne tulee vaan joku data virhe käy siis joku tämmöinen mitä nyt järjestelmissä aina voi käydä. Totta kai sitten voi olla että tulee virheitä, mutta siinäkin kai tärkeintä sitten se, että ne oikaistaan ja kertoa yleisölle mitä kävi, että minkä takia uutisissa me luki hetken aikaa, että Suomi voitti jonkun MM-jalkapallon kisat vaikka ei voittanut ja sitten ihmiset on pettyneenä siellä suihkulähteessä. [laughs]

Tähän liittyen puhuit niistä mustista laatikoista, niin jos on tämmöinen kone niin sehän voi vaikuttaa siihen uutiseen, jos ei tasan tiedä mitä se tekee. Miten tämmöisiä

toimintoja kehitetään? Kehitetäänkö näitä STT:n sisällä vai hankitaanko niitä ostopalveluna vai miten?

Meillä ei ole sisäistä siis tämmöistä koodariosastoa että ollaan sen verran pieni talo, että meillä ei tällä hetkellä ole sellaisia resursseja, että me pystyttäisiin in house tekemään vaikka joku NLP-malli ja sen takia meillä onkin siis otettu sitten osaa tämmöisiin yhteisprojekteihin, että on ollut esimerkiksi tämä Google rahoitettu, DNI-projekti joka oli Scoopmatic missä oli Turun yliopiston tutkijat ja nyt on sitten tää Embeddia, missä on tuota niin kuin EU-rahalla kansainvälinen tämmöinen konsortio. Että siis kyllä siinä yhteistyötä kannattaa tehdä ja sitten toki myös siis suomalaisen median kanssa. Sikäli kun aina löytyy semmoisia yhteisiä kiinnostuksen kohteita ja niitä missä tavallaan yhteiset edut ja kiinnostukset kohtaa, kannattaa tehdä sellaista yhteistyötä myös.

Onko teillä STT:llä ollut käytössä tämmöistä automaattista artikkelin versiointia tai sisällön summaamista, niin kuin ingressiin tekemistä?

Ei, ei ole toistaiseksi ollut.

Entä sitten miten näitä templateja, käytetäänkö niitä teillä, onko niitä tuotannossa nyt, tai miten niitä käytetään STT:llä?

Ei ole tällä hetkellä tuotannossa, että niissä on just se resurssiongelma mistä sanoin, eli tavallaan sitten jos on templateja pitää olla aina joku, joka niitä myös niin kuin huoltaa ja tuottaa ja tekee.

Aletaan olla loppusuoralla vielä muutama pikkuinen kysymys. Miten tuota STT:n hankkeisiin niin kuin liittymällä niin hakeeko STT kustannussäästöjä primäärisesti vai onko siinä myös ikään kuin muita faktoreita jotka tähän ajaa?

Siis tarkoitatko silleen tekemällä yhteistyötä vai hakemalla automaatiota?

Hakemalla automaatiota ylipäänsä.

No siis parhaassa tapauksessa kyllähän se voisi säästää myös kustannuksissa, mutta aivan varmaan ensisijaisesti siis se riittäisi meille tai olisi meille hyväksi avuksi, että jos se vapauttaa meidän ihmisten resurssia. Eli tavallaan vapauttaa sen meidän olemassa olevien toimittajien aikaa sitten johonkin mielekkäämpään. Siis vaikka nyt mietitään urheilutuloksia, että sitten jos kaikki niiden työstämiseen menevä aika, tai vaikka kaikki näiden talouden pörssikatsausten työstämiseen menevä aika vapautuisi sitten johonkin muuhun, vaikka sitten siihen, että sitä samaa aihetta pystyisi ihminen pohtimaan ja käsittelemään vähän syvällisemmin, niin se olisi sitten jo niin kuin parempaa palvelua sekä meidän asiakkaille, että sitten sille

yleisölle, jolle toki meidänkin uutiset viime kädessä tehdään. Että meillä ei ole siis semmoista ajatusta, että automaatio korvaa meidän toimittajia, vaan että jos miettii vaikka sitä Pikkulintua eli siis tätä seuranta-selain-osaa mistä aikaisemmin puhuin niin niin sehän on myös tehty nimenomaan siihen, että jotta meidän päivystäjät voi niin kuin rauhallisin mielin sitten tehdä kaikkea muuta kun ne tietää, että tietyt sivut on jo sen Pikkulinnun tarkkailussa, niin sitten ei tarvitse singahdella ja jakaa sitä huomiota kahteensataan paikkaan vaan sataviisikymmentä riittää. Eli tällaista haetaan siinä. Ja sitten toinen on ihan siis semmoinen hyöty sille meidän media asiakkaalle, eli että voitaisiin tavallaan tarjota semmoista sisältöä, mitä me ei tällä hetkellä tarjota. Esimerkiksi sitten vaikka niitä automatisoituja otteluselosteita vaikka jostain maakunnallisista sarjoista, joita me ei tällä hetkellä ei pystytä kaikista peleistä tekemään juttua, niin sehän olisi tavallaan hyvää palvelua sinne maakuntiin sekä medioille että lukijoille.

Joo tosi hyvä. Hyvä vastaus. No miten toimittajat reagoi tämmöisiin automaatioon liittyviin teknisiin kehityksiin.

Mulle on monta kertaa sanottu, että me ollaan - siis toimittajat - että me ollaan hyvin muutosvastarintaisia ja epäileväisiä. Mä en ole sitä mieltä. Mä en tiedä että johtuuko se siitä, että kun mä oon uutistoimistossa töissä, niin meillä on vaan niin paljon sellaisia työnkulkuja, johon tällainen sopisi tosi hyvin, että ihmiset on yleensä silleen, että jos niille sanoo että ”tiesittekö että jossain käytetään vaikka templateja siihen, että nämä talousluvut tulee suurin piirtein valmiina siihen toimitusjärjestelmään” ihmiset on vaan pelkästään innoissaan, koska he ymmärtää ihan yhtä hyvin että sitä työtä on vielä edelleen niin paljon, että se ei yksin riitä että ne talousluvut tulee että sitä ei voi silleen laittaa hattua naulaan ja lähtee loppupäiväksi lounaalle vaan sitten lähinnä miettiä, että mitä kaikkea mä voisin silläkin viidellä minuutilla tehdä muuta että se auttaisi. Semmoinen siis semmoinen varautuneisuus kyllä on, että siis ihmiset ymmärtää ihan siis tosi hyvin oman työnsä ja sitten niiden työnkulkujen realiteetit. Että tavallaan se, että jos tuodaan jotain uutta toimitukseen niin jos sitä vastarintaa on, niin mun oma käsitys tai semmoinen kokemus on se, että sitä on silloin jos ihmiset kokee että he ei saa tarpeeksi perehdytystä siihen asiaan, että sitä esimerkiksi pitää alkaa käyttää mutta et osaa, niin silloinhan se tuntuu siltä, että tämä vaikeuttaa mun työtä, koska joudun tappelemaan paitsi niiden talouslukujen kanssa, niin sitten sen templaten tai mikä se onkaan niin sen kanssa joka ne tekee. Eli tavallaan kaivataan semmoista perehdytystä. Sitten kaivataan perusteluja ja siis semmoista hyvin konkreettista, että miksi tämä auttaa juuri minua. Ja sitten semmoista pitkäjänteisyyttä. Eli siis silleen, että ne kanssa sitten jos jotain uutta tekniikkaa tuodaan niin siitä että sitä jaksetaan pitää pöydällä ja perehdyttää. Ja sitten myös siis semmoista aikaa, aikaa opetella ettei se sitten tavallaan tule silleen että yhden

vuoron alussa vaan saat kuvalliset ohjeet ja pärjäile. Että ikään kuin siis mun mielestä siihen tekniikkaan ei suhtauduta millään tavalla silleen periaatteellisesti torjuvasti, mutta ihmiset vaan ymmärtää sen, että jos kauheasti uusia asioita pitää kaiken muun oman työn ohella, otona tehdä niin kyllähän se sitten taas ruuhkauttaa sitä muuta tekemistä, että siitä mun mielestä vastustus tulee, mutta ei siitä, että että me oltais jotenkin ammattikuntana aina hirveän muutosvastarintaisia jääriä. Mut mä voin olla yksin tämän mielipiteeni kanssa. Moni tutkimus osoittaa muuta.

No toi kuulostaa jotenkin myös aika loogiselta, että se on enemmänkin näin kuin kuvailet kuin että ihmiset vastustaisi kauheasti työkaluja, jotka on tehty helpottamaan omaa työtä.

No jos toimittajat reagoi tälleen, niin miten yleisö reagoi ja mitataanko sitä?

Niin, uutistoimistolla on aina vähän semmoinen hitaampi, ehkä ikkuna siihen miten yleisö reagoi, kun julkaisijalla kun meillä ei ole sitä omaa julkaisukanavaa, niin me ei esimerkiksi sitten nähtäisi sitä, että kuinka paljon vaikka luetaan niitä automaattijuttuja. Tai että meillä ei olisi suoraan esimerkiksi näköyhteyttä sitten juuri niihin tiettyihin kommenttikenttiin, mihin ihmiset saattaisi jättää niiden ensimmäisen selkärankareaktion sitten sinne. Niin, sitä on vaikea sanoa, että miten yleisö reagoisi. Se varmaan riippuu ihan siitä, että mihin sitä käytettäisiin sitä automaatiota, kuinka avoimesti siitä kerrotaan ja kuinka laadukasta se sitten on. Et tavallaan, jos se on kovin laadutonta niin sittenhän se tuppaa ärsyttämään. Se on ihan sama onko se ihmisen vai koneen tekemä sisältö, mutta että tuppaa tavallaan tökkimään silmään. Mutta että jos sitten se on niin kuin, jos sen kokee hyödylliseksi niin en tiedä miksi tavallaan yleisökään suhtautuisi siihen erityisen torjuvasti, jos ne vaan tietää mistä on kyse. Eli siis tavallaan jos siinä ollaan läpinäkyviä eikä niin että luet kolme viikkoa jotain tosi mielenkiintoista palstaa vaan sitten neljännellä viikolla tajutakseni, että eihän tämä siis tämä, että tämähän on automaattista, että sitten siinä voi tuntea itsensä vähän jotenkin petetyksi.

Joo tosi hyvä.

Miten sä uskot että automaattinen uutistuotanto tai tämä ala tulee kehittymään viiden vuoden aikana? Tai kymmenen, tulevaisuudessa?

No siis mä uskon että niitä otetaan niitä sovelluksia entistä enemmän käyttöön. Siis ihan uutistoimituksessa, jos niistä vaan saadaan kustannustehokkaita tietyllä tavalla siis sekä myös aika tehokkaita ja käyttäjäystävällisiä ja sitten luotettavia, että tavallaan että jos ne niin kun tekee hyvää journalismia ja tällä tarkoitan siis sitä että ne esimerkiksi pysyy faktoissa, eikä hallusinoi niin kyllähän

niitä sitten otetaan käyttöön. Mutta että sitten heti, jos siellä on vaikka jotain tämmöisiä luotettavuusongelmia, tai vaikka jotain isompia tietoturvaongelmia. Tai niin sitten voi olla, että tavallaan tunnistetaan kyllä se tarve ja se hyöty mitä siitä voisi olla. Mutta sitten taas että kyllä niin kuin varmaan journalismissa esimerkiksi luotettavuus on kyllä ihan viimeinen, joka sitten uhrataan sen eteen, tai toivon, että näin on, sen eteen, että on helpompi tehdä jollain koneella. Jos laatu on riittävää niin tänne vaan. Mutta sitten jos tavallaan ihminen joutuu joka tapauksessa kaiken oikolukemaan, niin sitten voi samantien tehdä itse.

Hienoa, se oli mun viimeinen kysymys. Onko sinulla jotain, mitä sinä itse haluaisit sanoa tai pidät tärkeänä? Niinku mitä ottaa huomioon?

Eipä oikeastaan siis tosi mielenkiintoinen aihe, että jäämme mielenkiinnolla odottamaan, että mitä löydöksiä tulee